PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

ADMISSION PLANNING IN HEALTHCARE: STOCHASTIC AND HIERARCHICAL APPROACHES

ANA BATISTA GERMÁN

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Advisors:
JORGE VERA
DAVID POZO

Santiago de Chile, December 2020

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

ADMISSION PLANNING IN HEALTHCARE: STOCHASTIC AND HIERARCHICAL APPROACHES

ANA BATISTA GERMÁN

Members of the Committee:

JORGE VERA
DAVID POZO
SERGIO MATURANA
ENZO SAUMA
SUSANA MONDSCHEIN
PHILIPPE CHEVALIER
VLADIMIR MARIANOV

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences

Santiago de Chile, December 2020

To my parents José y Mariana.

ACKNOWLEDGEMENTS

Pursuing a Ph.D. has been one of my greatest challenges. Staying far from home and family in a completely new environment where I had to start a new life with different standards was not easy. It would have been impossible to complete this journey if it was not for the help of God first and the support of a number of people I would like to express my deepest gratitude to.

I want to start by expressing gratefulness to my supervisor Dr. Jorge Vera for believing in me from the moment I applied for the doctorate program. He taught me to work independently and to develop my researching abilities. I feel very honored for being a part of his team and receiving his knowledge and unconditional support during the entire process.

I am also grateful to my second supervisor Dr. David Pozo. It would not be possible to take this thesis to its completion if it was not for his guidance. There is not enough place to express how thankful I am for his support, both intellectually and personally. David's energy and enthusiasm about investigating have motivated me to keep on developing this thesis even through the difficult stages of this process. He is a great mentor who showed me how to see problems from a different angle. I want to thank him not only for being my professor but for being my friend as well. I will always remember humbleness and generosity among his traits.

I would like to acknowledge the research committee composed of Dr. Sergio Maturana, Dr. Enzo Sauma, Dr. Susana Mondschein, Dr. Philippe Chevalier, and Dr. Vladimir Marianov for their remarks that were of great use for improving my thesis. I would also like to thank Dr. Van-Anh Truong at Columbia University for the valuable insights on topics related to this thesis.

I am thankful to the administrative team of the school of engineering of the Pontifical Catholic University of Chile, to the secretaries and administrative support, for their help

at crucial stages during my doctorate studies. Special thanks to Fernanda Kattan, Nicole Betti, Pilar Martinez, and Karina Norambuena.

I want to thanks the National Commission for Scientific and Technological Research (CONICYT) in Chile for providing me financial support to carry out this doctorate thesis through grant 6635/014. Thanks to this grant, I was able to participate in several national and international conferences, which helped me to enrich my investigation. It also gave me the opportunity to have a doctorate internship at Columbia University, New York. I would also want to thanks the medical and administrative team of the hospital we work with in Chile, who provided me valuable insights to understand the Chilean health care system and its operations.

My friends and colleagues have been an inalienable part of this journey; they have always been around to cheer me up during difficult days. I would like to give special thanks to Jheser, Cesar, and Armin, who were not only my friends but also my "personal coaches." I am thankful to them for always being willing to help in the critical moments of this process. I am very grateful to Jheser for encouraging me to start my research career in Chile; his insights helped me from the beginning to establish myself in this new journey. I have learned to be persistent and disciplined from Cesar, who was so generous, and thanks to his family, he is also a great dancer. Armin taught me to control my emotions and concentrate on the most important things; vital advice to achieve finish a Ph.D. I would also like to say thank you to all the other friends and colleagues whom I met throughout these years and who have a special place in my heart; my wise friends and colleagues from the Pontifical Catholic University of Chile, specially to the joyful Latin crew in the engineering department, my beloved brothers and sisters in the faith of the Neocatechumenal way community, the remarkable friends from the Maria Luisa Santander residence, my beautiful and cheerful Venezuelan friends, the charismatic colleagues from the Columbia University in New York, and my colleagues and friends from the Skolkovo Institute of Science and Technology in Moscow.

I want not only to express my gratefulness but to dedicate this thesis to my family and especially to my parents José and Mariana. The Christian religious education that I received from them was essential during this time. Believing in God looking for his help has supported me in this path. I am grateful to my brothers and sisters, brothers-in-law and sisters-in-law, nephews and nieces who, despite the distance, could make my day brighter with a simple *"hello, how are you"*. I feel happy to have them all in my family. Last but not least, I am deeply grateful to Sergio, who has been crucial support in the last stage of this journey, helping me in a country that I would never imagine myself to live in - Russia. I want to express to him my gratitude for teaching me patience, being less self-centered, and making me get to know more about myself.

I would like to close this section with a self-awareness quote. The path to achieving this Ph.D. has been a remarkable journey where, besides professional development, I became a better version of myself. I had extreme ups and downs, made incredible friends with whom I shared tears of sadness and joy, and experienced charming places and moments around the globe I will talk about for a lifetime. I feel deeply grateful for all of that, and I hope my work can inspire others. *That's a wrap!*

Contents

List of Tables

List of Figures

List of Acronyms

ADP	Approximate Dynamic Programming.
AHA	American Hospital Association.
APP	Admission Planning Problem.
CAD	Central Admission Department.
CCG	Column and Constraint Generation.
CS	Clinical Service.
CU	Critical Unit.
DP	Dynamic Programming.
DRCC	Distributionally Robust Chance Constrained.
DRG	Diagnosis Related Group.
DRO	Distributionally Robust Optimization.
ED	Emergency Department.
ER	Emergency Room.
EVPI	Expected Value of Perfect Information.
GDP	Gross Domestic Product.
GoF	Goodness-of-Fit.
ICU	Intensive Care Unit.
ILP	Integer Linear programming.
KL	Kullback-Leibler.
LoS	Length of Stay.
MDP	Markov Decision Process.
MILP	Mixed-integer linear programming.
MIP	Mixed-integer programming.
NE	Network allocation.

OECD	Organisation for Economic Co-operation and Development.
OR	Operating Room.
OU	Over-utilization.
PACU	Post-Anesthesia Care Unit.
PBAP	Patient Bed Assignment Problem.
RE	Recovery room.
RI	Reliability Index.
RO	Robust Optimization.
SAA	Sample Average Approximation.
SII	Severity Illness Index.
SOCP	Second-Order Cone Program.
SP	Stochastic Programming.
SU	Surgery room.
TSO	Two-stage Stochastic Optimization.
UU	Under-utilization.
VSS	Value of Stochastic Solution.

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

ADMISSION PLANNING IN HEALTHCARE:
STOCHASTIC AND HIERARCHICAL APPROACHES

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the Degree of Doctor in Engineering Sciences by

ANA BATISTA GERMÁN

ABSTRACT

The healthcare system is a service industry that requires quality planning due to the relevance in its operations. In particular, public hospitals face constant pressure to be cost-efficient. A central process to help manages the hospital's operations is admission planning, which aims to ensure timely access and efficient use of resources. However, several sources of uncertainty and resource limitations challenge the accomplishment of those objectives. Besides, similar to many industries, the hospital's long-term decisions are based on aggregated plans, which should be fulfilled in the short term, subject to uncertainty. Since no coordination between temporal levels is guaranteed, several inconsistencies may arise, such as unnecessary waiting times, rejections, and even early death of patients.

This thesis focuses on assessing the problem of temporal consistency in the admission planning problem, which is subject to uncertainty and constrained for bed capacity. The problem dwells on how to develop a tactical plan that considers the impact of operational decisions while guaranteeing feasibility. The main objective is to model and solve a hierarchical decision-making process that integrates the tactical and operational levels of planning to enhance service quality and efficient use of resources.

The research conducted in this thesis contributes to providing efficient decision frameworks to solve the admission planning problem. We consider optimization methods under uncertainty in a multi-stage fashion to improve consistency and coordination of planning. A bi-objective two-stage stochastic approach is developed to study allocation decisions under demand and bed availability uncertainty and assess the trade-off between patient and hospital perspectives. Since the access to complete information is often limited, an adaptive distributionally robust optimization approach is proposed to study allocation and scheduling decisions. Partial distributional information of the patient length of stay is considered and modeled through an enhanced formulation for service-time-type constraints. The approach allows evaluating the balance between robustness and consistency.

Through extensive numerical studies and validation using real data, we show that efficient planning is obtained by adopting a hierarchical stochastic framework of decision. The framework minimizes the substantial inconsistencies involved in admission planning. Managerial insights are provided for tactical-operational planning.

Keywords: hierarchical planning, intertemporal planning, admission planning, uncertainty, robustness, healthcare, capacity planning.

Members of the Doctoral Thesis Committee:

JORGE VERA
DAVID POZO
SERGIO MATURANA
ENZO SAUMA
SUSANA MONDSCHEIN
PHILIPPE CHEVALIER
VLADIMIR MARIANOV

Santiago de Chile, December 2020

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
ESCUELA DE INGENIERIA

PLANIFICACIÓN DE ADMISIÓN EN SALUD:
ENFOQUES ESTOCÁSTICOS Y JERÁRQUICOS

Tesis enviada a la Dirección de Postgrado en cumplimiento parcial de los requisitos para el grado de Doctor en Ciencias de la Ingeniería.

ANA BATISTA GERMÁN

RESUMEN

El sistema de salud es una industria de servicios que requiere planificación de calidad debido a la relevancia en sus operaciones. En particular, los hospitales públicos enfrentan presión constante para ser eficientes. Un proceso clave que gestiona las operaciones hospitalarias es la planificación de admisiones, que tiene como objetivo garantizar acceso oportuno y uso eficiente de los recursos. Sin embargo, varias fuentes de incertidumbre y limitaciones de recursos desafían el logro de esos objetivos. Además, similar a muchas industrias, las decisiones de largo plazo del hospital se basan en planes agregados que deben cumplirse en el corto plazo, el cual está sujeto a incertidumbre. Dado que no se garantiza coordinación temporal, pueden surgir varias inconsistencias, como tiempos de espera innecesarios, rechazos e incluso muerte prematura de los pacientes.

Esta tesis evalúa el problema de consistencia temporal en el problema de planificación de admisiones, el cual está sujeto a incertidumbre y limitado por la capacidad de camas. El problema radica en cómo desarrollar un plan táctico que considere el impacto de las decisiones operativas mientras que garantice su viabilidad. El objetivo principal es modelar y resolver un proceso jerárquico de toma de decisiones que integre los niveles táctico y operativo para mejorar la calidad del servicio y el uso eficiente de los recursos.

La investigación contribuye a proporcionar marcos de decisión efectivos para resolver el problema de planificación de admisiones. Consideramos métodos de optimización bajo incertidumbre en múltiples etapas para mejorar la consistencia y coordinación de la planificación. Se desarrolla un enfoque bi-objetivo estocástico en dos etapas para estudiar las decisiones de asignación, considerando incertidumbre de la demanda y disponibilidad de camas y evaluar el *trade-off* entre la perspectiva del paciente y la del hospital. Dado que el acceso a información es limitado, también se propone un enfoque de optimización robusta distribucional adaptativa para estudiar decisiones de asignación y programación. Se considera información parcial distribucional del tiempo de estadía, modelado a través de una formulación mejorada para restricciones de tipo de tiempo de servicio. El enfoque permite evaluar la relación entre robustez y consistencia.

Mediante extensos estudios numéricos y validaciones con datos reales, demostramos que se obtiene una planificación eficiente adoptando un marco de decisión estocástico jerárquico. El marco es capaz de minimizar las inconsistencias implicadas en la planificación de admisión. Se proporcionan lineamientos de gestión para planificación táctica-operativa.

Palabras Claves: planeación jerárquica, planeación intertemporal, planeación de admisión, incertidumbre, robustez, salud, planeación de capacidad.

Miembros de la Comisión de Tesis Doctoral:

JORGE VERA
DAVID POZO
SERGIO MATURANA
ENZO SAUMA
SUSANA MONDSCHEIN
PHILIPPE CHEVALIER
VLADIMIR MARIANOV

Santiago de Chile, December 2020

Chapter 1. INTRODUCTION

This thesis's primary purpose is to study the problem of temporal consistency in decision-making for admission planning under bed capacity limitations. We aim to model and solve a hierarchical decision-making process at the tactical-operational levels to guarantee the expected levels of service and resource utilization.

Section 1.1 provides an overview of the healthcare system, its main characteristics, and its concepts. This section also introduces the admission planning problem within healthcare capacity planning, which is the focus of this thesis. Section 1.2 describes intertemporal planning decisions. Section 1.4 presents the thesis motivation and scope of study. Section 1.5 provides state of the art. Section 1.6 states the thesis hypothesis and objectives. Finally, Section 1.7 presents the structure of the thesis, a summary of each chapter's content, and the contributions.

1.1. Healthcare system: Overview

The healthcare system is a socioeconomic service industry that comprises complex and costly operations. During the past decades, health care providers have faced government pressure to improve efficiency and quality of care due to the rising expenditures. In Chile, for instance, health costs accounted for 6.7% of the Gross Domestic Product (GDP) in 2010; in 2018, it increased to 8.9%. In the United States, the increase has been even more relevant; It accounted for over 14% of the GDP in 2018 compared to 8% in 2010 (OECD, 2020), and it is expected that these rates continue to boost. The aging population, the increase of chronic care demand, and the gap between technology development and efficiency improvement are the leading causes of the rising expenditures, in a system with limited resources and uncertainty (Reid & Grossman,

2005). In addition, as mentioned in Brailsford and Vissers (2011) the service concept in healthcare has changed from focusing only on productivity improvement to find a balance between service quality and efficiency. Thus, it is crucial to ensure the quality of care achieved through efficient planning and control of resources.

The structure of the healthcare system varies among countries according to particular needs and provision of resources. For instance, countries with market competition such as the United States are focused on service improvement; on the contrary, countries based on budgeting seek to improve resource efficiencies, such as most European countries, Chile, Canada, and Australia (Brailsford & Vissers, 2011).

Independently of its structure, according to Fanjiang et al. (2005), the healthcare system is composed of four interrelated elements:

(i) the patient,
(ii) the care team (e.g., care providers, physicians, family),
(iii) the organization (e.g., hospitals, clinics, nursing homes), and,
(iv) the environment (e.g., regulators, insurers).

The improvements implemented in each element will have a significant impact on the service level offered to patients, and the efficiency in the use of resources. In particular, *hospitals* are a critical area because it is where the highest healthcare system expenses are generated.

The hospitals provide the infrastructure for healthcare delivery. It comprises the decision-making systems, information systems, operating systems, and processes needed to provide quality care (Reid & Grossman, 2005). Depending on the source of funding, hospitals can be classified into public and private institutions; the former is publicly funded by the government, while the second recovers its costs through service-based payments schemes such as private insurance companies and self-pay (Vijayakumar et al., 2013).

Thus, public hospitals have to manage the system operations depending on the annual budget limits determined by the government.

The management of hospital operations comprises several managerial areas: medical planning, materials planning, financial planning, and resource capacity planning (Hans et al., 2012). Medical planning is related to decisions taken by clinicians, such as medical protocols and diagnosis planning. Material planning refers to the acquisition of goods needed to perform health care. Financial planning addresses the management of costs and revenues, from investment decisions to the billing process. Finally, resource capacity planning concerns decisions from capacity dimensioning to resource scheduling, e.g., workforce, equipment, and facilities.

From the capacity planning perspective, a hospital can be thought of as a network of interrelated processes and services whose capacity is finite, and patients are the main flow and source of resource consumption. The authors in Gemmel and Van Dierdonck (1999) compare a hospital to a manufacturing system that is conformed by heterogeneous jobs, processes that are initiated by the customer and, the throughput times are short. Additionally, the demand (i.e., the patient) is time-dependent, the production and consumption have to be performed simultaneously, and the unit to be produced is decomposable. Despite the similarities with other industries or commercial services, hospital management is different because human beings are the dominant factor of improvement. The *care* is the product of the system, and the health professionals are the producers (Rauner & Vissers, 2003). Unlike manufacturing industries, care is not a commodity that can be stored. The mismatch between capacity and demand can only be managed with the use of slack capacities or buffers of patients before being admitted to the hospital (i.e., waiting lists) or after being admitted (i.e., waiting times) (J. M. Vissers et al., 2001).

The resource capacity planning in hospitals implies several processes, including staff planning, equipment dimensioning, and admission control. Of particular importance is the admission control process as it is a central process that impacts hospital operations and, therefore, costs and profit. As stated in the taxonomy of Hulshof et al. (2012), the admission control process tackles decisions at different care services and planning levels. In the next section, we describe the admission process as a dependency of capacity planning.

1.1.1. Admission control in healthcare systems

The *admission control* is a critical process within healthcare capacity planning that aims to ensure timely access and efficient use of resources. The *Admission Planning Problem* (APP) is a subtask of the admission control process, that seeks to smooth the workflow of patients and to reduce waiting times, delays, and cancellations. Coping with variability in patient arrival, patient's stay, and resource utilization is the main challenge.

Three main concepts drive the admission decisions in hospitals:

(i) *the type of care*: it accounts for the classification of care services, mainly subdivided into *outpatient* and *inpatient*. The former is related to services that do not require hospitalization but medical attention or treatment, such as ambulatory care, home care and, residential care. On the other hand, the latter accounts for services in which the patient needs hospitalization and will stay on a bed for a period according to their clinical condition, for instance, emergency care, surgical care, inpatient care,

(ii) *the level of care*: it defines the patient's acuteness, that is subdivided into primary, secondary, and tertiary. Primary care is the lowest level of acuity and is usually served in the outpatient service. The secondary and tertiary levels

of planning are for patients with pathologies that require hospitalization, and it concerns services in inpatient care, and finally,

(iii) *the available resources*: it varies regarding the type of care, for instance, doctor, nurses, specialized equipment (e.g., MRI, X-ray, scanner), facilities, operating rooms and, beds.

According to the type of care (i.e., outpatient, inpatient), the admission process accounts for different requirements. In the outpatient setting (i.e., primary care), the resources are mainly physicians and medical equipment. Also, most of the patients require fixed time length of services; thus, the care is divided into time slots of the same length. On the contrary, in the inpatient service (i.e., secondary and tertiary care), the patient service time, commonly known as the *Length of Stay* (LoS), tends to be random according to the patient's diagnosis and, therefore it affects the use of resources by causing fluctuations of capacity availability. For the inpatient service, the primary resources are bed capacity, medical staff, and operating room. Traditionally, bed capacity has been the main unit of planning in a hospital since it affects spending, quality of care, and patient accessibility (Green, 2002). Overall, the inpatient setting is more challenging to manage than the outpatient since it involves more complex patient care (Pierskalla & Brailer, 1994), and therefore, more data and efficient information systems are needed.

The admission planning in the inpatient setting generally considers two categories of patients, *unscheduled* and *scheduled*; the first one refers to the unplanned admissions and is commonly termed as *emergency* patients, while the latter concern the *elective* admissions of patients who have previously visited a specialist and need hospitalization. The admission is usually managed by a Central Admission Department (CAD), that receives requests for admission from different departments in the hospital, i.e., emergency, surgery, and other clinical services, to be allocated in the corresponded care units such as

Intensive Care Unit (ICU), intermediate unit, and nursing wards. The CAD also receives requests for admission from other hospitals in the healthcare network. The beds that belong to each area are categorized according to the level of diagnosis acuity; the ICU is for patients with a high acuity level, while the intermediate unit is a previous step to occupy the general nursing wards.

The general scheme of the admission process and patient flow in the inpatient service is illustrated in Figure 1.1, in which several flows can be identified. The unscheduled (i.e., emergency) patients get into the emergency room and could require surgery in the Operating Room (OR), and the use of the Post-Anesthesia Care Unit (PACU). The scheduled patients may need the OR before being allocated in the corresponding units or going directly to the nursing wards. Two main *storage* points or waiting areas (inverted triangle symbol) can be identified in the admission flow, indicating the list of patients waiting to be admitted to the hospital. The waiting list corresponds to requests for admission of different care units in the hospital. Note that there is no waiting list for unscheduled patients because those patients are supposed to be served as soon as possible. The first waiting list point (intersection between scheduled and operating theatre block) is for scheduled patients waiting for surgical intervention and, the second one is for patients waiting to be allocated in a nursing ward.

During the hospitalization and according to the patient's condition and the availability of beds, he/she may be transferred between care units. The process concludes with the patient discharge from the nursing ward, which can be due to a voluntary decision, medical referral, or by death. Additionally, a patient may be sent to an external location, i.e., a private or public hospital in the healthcare network. The hospital can also receive applications for admission from external hospitals. Figure 1.1 illustrates that overall, all of the admission flows require a nursing ward to be allocated. The wards correspond to

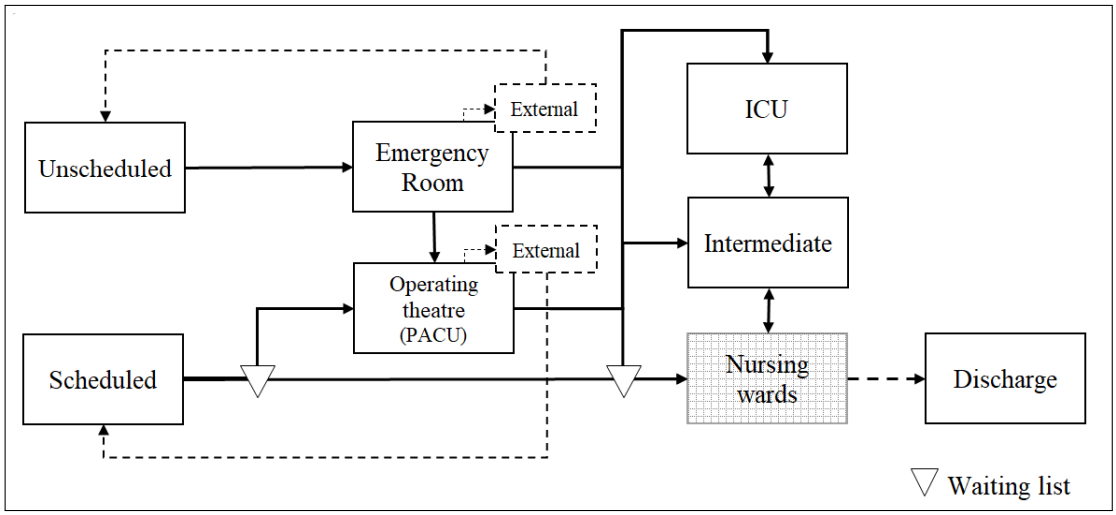


FIGURE 1.1. General scheme of the admission process and patient flow in hospitals.

the most significant mass of allocation in the hospital; therefore, it needs more attention in the admission process because it can cause bottlenecks in the inpatient flow.

As seen in the inpatient process flow, admission decisions are complex since they involve different interrelated processes. In general, the admission decision process can be classified into two phases: *allocation* and *scheduling*. Allocation decisions aim to decide which patient to admit considering resource availability. Scheduling decisions address the problem of determining the time and date of admission for assigning patients to beds or rooms for the duration of their stay. Such decisions, similar to manufacturing and other service settings, could be described according to a *temporal hierarchy*.

In the capacity planning and control framework applied to healthcare presented in Hulshof et al. (2012), the authors considered the well-known hierarchical decomposition of decisions as defined in Anthony (1965); strategic, tactical and operational, but subdividing the operational level in *offline* and *online*. The hierarchical decisions are described as follows:

(i) Strategic planning: This level considers a long horizon of multi-year planning at an aggregated level. The admission decisions are related to service mix, case mix, and capacity dimensioning of the hospital services. The performance criteria involve the setting of service levels per patient group (i.e., waiting time, cancellation rate, deferred admission), and the definition of resource utilization targets (i.e., bed occupancy).

(ii) Tactical planning: The planning at this level considers mid-term decisions on a monthly to a weekly horizon. It concerns decisions of determining the mix of patients, i.e., *which* patient to admit and the definition of admission policies for patient groups. It also determines capacity reserve rules for unscheduled patients (i.e., emergency patients). In addition, this level of planning involves decisions about *where* (i.e., bed, room/ward) to allocate the patient, by considering several features such as hospital resource availability, medical requirements, patient needs, among others.

The problem of bed allocation is a subproblem of the APP that has been addressed in the technical literature as Patient Bed Assignment (PBAP). At the tactical level, the PBAP defines aggregated scheduling policies of patient groups to be allocated in rooms (i.e., patient-to-room). The performance measures are related to the evaluation of the criteria previously established at the strategic level, such as waiting time and bed utilization. The PBAP further considers measures of patient throughput.

(iii) Offline-operational planning: it focuses on short-term decision-making in days and up to a couple of weeks. The decisions stand for *in advance* planning, usually related to decisions of scheduling individual patients, (*who*), to provide dates and times (*when*), of admission in a more detailed fashion. At this level, the PBAP determines scheduling policies and time allowances for patients

to be assigned in a specific bed/room. The performance measures consist of evaluating resource under-utilization and over-utilization metrics due to fluctuations in bed utilization. Also, it evaluates measures of overstay and overtime.

(iv) Online-operational planning: This level of planning refers to short-term decisions on the day to hours of patient admission at a detailed level. The decisions are made reactively to control unforeseen events such as the arrival of emergency patients. It determines buffers for handling daily fluctuations and the definition of priority rules of admission. The performance criteria are similar to offline operational planning but are evaluated dynamically.

Decisions at the strategic level in public hospitals are defined by governmental institutions that decide capacity levels and targets of resource utilization; thus, at this level, limited flexibility can be achieved. The operational level also has low responsiveness since the decisions have to be taken reactively, given that the upper levels decisions impose constraints to the lower level planning. Then, in comparison to the strategic and operational levels, the tactical planning, which lies in between, allows more *flexibility*, focusing on operations/execution decisions. For example, at the strategic and operational levels, capacity requirements are fixed; however, at the tactical level, temporary capacity expansions, such as overtime or additional resources, could be added (Hans et al., 2012). Improvements in terms of cost-efficiency are thus better implemented at the tactical-operational (offline) levels of planning. Figure 1.2, summarizes the functions of each decision level for the inpatient service's admission planning problem.

From Figure 1.2, we observe that there is an evident interrelation between the temporal levels. The strategic level focuses on aggregate long-term decisions, such as sizing and case-mix planning, while the tactical-operational levels concern the execution

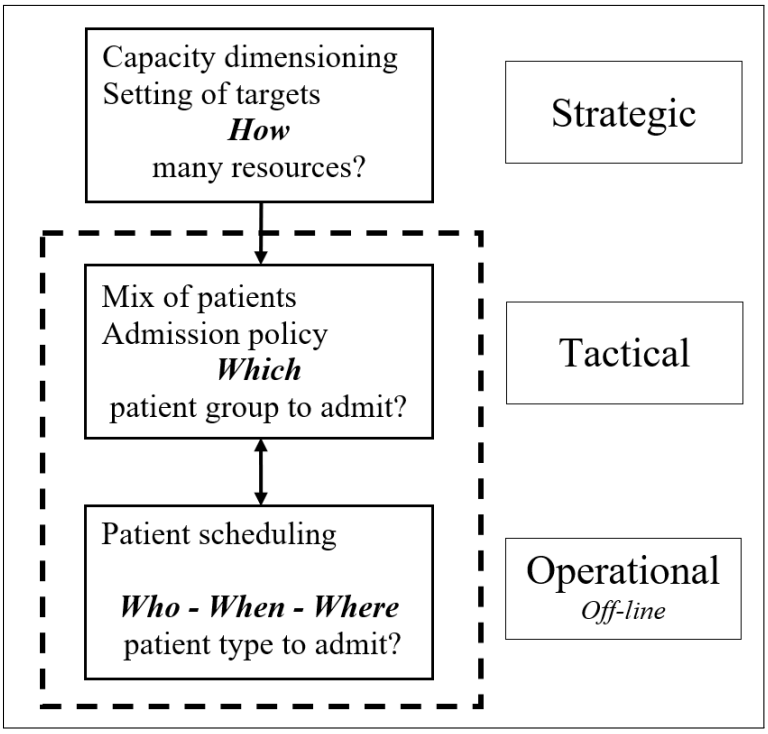


FIGURE 1.2. Temporal decomposition of the decisions in the APP.

of the processes. Since, at the strategic level, there is no detailed information available, decisions are made with aggregate information. Then, at the tactical-operational level, when the decisions are executed, the plan may result in inconsistencies. Such inconsistencies, e.g., cancellations, patient diversion, waiting time, overstay, affect not only the quality of patient care but also the performance in the use of resources. Therefore, the admission plan must be executed, guaranteeing coordination between the different temporal levels.

1.2. Intertemporal planning

The planning process in most industries is usually managed hierarchically. In practice, at upper levels, industries perform aggregate plans to anticipate demand and establish optimal production and inventory (Nam & Logendran, 1992). Then, in the short term, it is

necessary to perform scheduling tasks. As the aggregation level decreases, the aggregate plan may not be directly feasible due to variability and uncertainty. Several issues may arise if the decisions at each level are carried out independently because higher levels impose constraints on lower-level actions. In other words, lower level decisions can be used to reevaluate the decisions taken at upper levels (Carravilla & de Sousa, 1995). Therefore, there exists a need for coordination between the different levels of planning to achieve a certain level of consistency and feasibility. This decision-making process is known as *hierarchical planning*.

The hierarchical planning concept was first introduced by Hax and Meal (1975) in the production and planning setting. The approach divides a set of problems in subproblems to be solved sequentially. This method aims to achieve coordination between decision levels, i.e., long-term aggregate planning with short-term disaggregate planning. Three main concepts have been employed in the literature to define a hierarchical decision process (Yan, 2000), namely, (i) product disaggregation that divides the production plan into product types, families, and items to solve each problem sequentially (Bitran & Hax, 1977; Kira et al., 1997); (ii) process decomposition which is commonly applied in the manufacturing setting divides the decision-making in sequential processes (Yan, Xia, Zhu, Liu, & Guo, 2003), and (iii) temporal decomposition that consists of the decomposition of the decisions within the planning horizon (i.e., strategic, tactical, operational), also known as *intertemporal planning* (Lobos & Vera, 2016; Alvarez et al., 2020).

The application of the concepts above within a hierarchical decision approach has shown several advantages in decision-making. According to Dempster et al. (1981), there are two main advantages of using a hierarchical approach methodology:

(i) *the reduction of complexity* because it allows ensuring interaction between subproblems.

(ii) *the management of uncertainty* given that it postpone the detailed decisions at lower levels to guarantee that decisions made at higher levels are consistent at a certain cost.

Although a hierarchical approach provides several benefits, its application is prone to cause, inconsistencies, infeasibilities, and suboptimality if it is not effectively implemented (Beaudoin et al., 2008). The definition of such concepts is clearly explained in Beaudoin et al. (2008); inconsistencies are the result of conflicting objectives at different planning levels, infeasibility occurs due to information aggregation, and suboptimality refers to the lack of quality interactions between decision levels.

Therefore, in order to achieve efficient planning, it is necessary to consider the adequate methodology able to acknowledge coordination between decision levels for conflicting objectives, reduce complexity, and that incorporates uncertainty for different levels of disaggregation in the planning process. Optimization methods under uncertainty within an intertemporal approach of decisions may provide the proper framework to develop efficient planning.

1.3. Optimization frameworks for decision making under uncertainty

Optimization methods under uncertainty at different decision stages linked in time have been shown to provide efficient planning (Bertsimas & Goyal, 2010). Accordingly, within such a decision framework, strategic and tactical planning performed with data aggregation under uncertain conditions and limited information at a lower level may reduce the infeasibilities and inconsistencies typically found in the plan implementation.

Several optimization methodologies can be adopted. The most common being Stochastic Programming (SP) and Robust Optimization (RO). Within stochastic programming, Two-stage Stochastic Optimization (TSO) is a type of integrated approach

in which decisions are divided into two stages (Birge & Louveaux, 2011). The first stage is an aggregate level of decision and is taken before the realization of uncertainty. At this stage, the impact of decisions at lower levels have to be considered. The second stage captures the variations related to lower-level decisions, subject to uncertainty. In two-stage optimization models, uncertain parameters are modeled as random variables, and the probability distribution is assumed to be known.

In contrast to two-stage stochastic optimization, RO does not require detailed probabilistic knowledge of the uncertain information. Such a feature allows the problem's solution to be independent of data variability, modeled within an uncertainty set (Bertsimas et al., 2011). The decision-making can be sub-divided into a two-stage or multi-stage framework, in which the decision variables can be adapted in response to uncertain realizations (Ben-Tal et al., 2004).

Another methodology to study intertemporal problems is Distributionally Robust Optimization (DRO), which can be seen as a generalization of two-stage optimization and classical robust optimization (Wiesemann et al., 2014). In contrast to TSO and RO, distributionally robust optimization captures the decision-makers ambiguity aversion while assuming partial knowledge of the distributional characteristic of the uncertain parameters.

The optimization methodologies mentioned above share several similarities and differences, as well as advantages and disadvantages in its implementation. In Chapter 2, we explain in detail the theory of decision-making under uncertainty, including TSO, RO, and DRO methodologies.

1.4. Thesis motivation

The hospitals face constant pressure to be cost-efficient. In particular, public hospitals are under governmental control on the expenditure and must guarantee high levels of service and the fulfillment of resource utilization targets in a system with limited resources. The improvement in resource utilization in hospitals plays a crucial role in determining capacity decisions (e.g., medical staff, beds, equipment). Of particular importance in hospital capacity planning is the *admission planning process* that aims to ensure timely access and efficient use of resources. In this process, *the beds* are a critical resource since they represent the place where the patients are allocated for their stay, and thus it serves as a measure of hospital capacity (Green, 2002).

The decisions about the accomplishment of the level of service and resource utilization are conflicting objectives. For instance, hospitals aim to obtain high bed occupancy levels while patients expect to be treated with good service quality. However, this objective is diminished when capacity levels are high since it restricts the admission of new patients. Achieving these goals is particularly challenging because *care* is the main factor in the delivery of health services, causing uncertainty in the decision-making process. Additionally, due to the technological advances and the increase in acute care demand, the inpatient admissions are growing globally and, therefore, the management costs. In the United States, for instance, the number of admissions in 2019 increased by one million patients compared to 2018 (American Hospital Association [AHA], 2019), which represented an expense of over $1 trillion in admitted patients. Despite this, according to the Organisation for Economic Co-operation and Development (OECD), the number of beds has dropped in almost all countries, on average, from 5.5% per 1000 inhabitants in 2000 to 2.5% in 2018 (OECD, 2020). Thus, the mismatch between demand

and capacity, along with the system uncertainty, complicates decision-making in service delivery.

Several drawbacks can be identified as a result of the inefficiencies in the admission process, such as cancellations, patient diversion, the use of temporal capacity, fluctuations in bed utilization and, prolonged waiting time, which by all means affect the level of service offered to patients as well as the health spending. For instance, in Chile, patients in the emergency service wait over five days to be admitted to a ward bed. Additionally, by 2017, more than 270,000 patients were held on the waiting list, of which most of them have been waiting for more than a year (Bedregal et al., 2017). Even more regrettable, in the first half of 2018, about 9000 people died while waiting on the waiting list, representing a 54% increase compared to 2017 (Rebolledo, 2019). Such trends suggest the need to develop robust admission plans to improve the hospital's performance through resource utilization and service level.

In order to perform an admission plan, it is necessary to have information about: (i) how much capacity is needed, (ii) how much capacity is available, and (iii) which policy to implement to achieve the expected goals (Groot, 1993; Gemmel & Van Dierdonck, 1999). However, hospitals usually lack relevant information to develop robust plans, mainly due to several sources of uncertainty revealed at the operational level and are difficult to estimate, such as bed availability, arrival pattern of unscheduled patients, and the patient's length of stay. Note that as stated in Joosten et al. (2009), such variations could be *artificial*, which must be eliminated, e.g., scheduling methods of medical appointments, and *natural*, which can be administered but not eliminated, e.g., individual differences between patient's diagnostic.

The management of patient admissions in most hospitals is reactive rather than proactive. The current practice is based on operational strategies over fixed decisions

established strategically, which may lead to suboptimal results. For instance, hospitals set targets of bed occupancy at the strategic level, expected to be achieved at the operational level. This policy may result in low capacity availability at the operational level to admit the uncertain demand from unscheduled patients. Note that while scheduled patients can be held in waiting lists (i,e., capacity buffers), the unscheduled patients have to be served immediately; hence, slack capacity is needed to avoid bed shortages and reduce delays, waiting time, and transfers. Another common strategy in hospitals to deal with the variations and uncertainty is to increase bed capacity. However, this strategic measure may result in performance deviations (i.e., under-utilization and over-utilization) if the necessary adjustments are not made at lower levels, i.e., tactical-operational.

The above considerations lead us to infer that, to derive robust admission plans and to ensure feasibility and coordination in the planning process, the plan carried out at the aggregate level needs to be consistent with the disaggregated plan in the short term, subject to uncertainty and variability.

In the past years, operational researchers have shown increasing interest in healthcare applications aiming to improve the quality of care and efficiency in its complex operations (Green, 2012). In particular, the APP with bed capacity constraints has been studied extensively in the technical literature (He et al., 2019; Teixeira & De Oliveira, 2015; Hulshof et al., 2012), due to its impact in hospital operations. The studies focus on minimizing patient overcrowding in the service units, the number of canceled interventions, and transfers due to the lack of beds. From the mathematical modeling point of view, the APP has the structure of the bin-packing model, which is known to be NP-hard (Korf, 2002). Due to its combinatorial complexity, most existing models are heuristics, and metaheuristic-based methods used approximation algorithms (see, e.g., Demeester et al. (2010); Ceschia and Schaerf (2011, 2012); Turhan and Bilgen (2017); Guido et al.

(2018)). Although the models capture interesting characteristics of real problems in size and structure, they neglect the guarantee of temporal consistency in the decision-making and the uncertain nature of the APP.

Other contributions focus on modeling the variability and uncertainty of parameters, such as patient arrival, length of stay, and capacity availability. The most common applied approaches are simulation and scenario-based frameworks (He et al., 2019), focused on a single level of decision (e.g., strategic and operational). Other studies consider analytical approaches such as queuing theory (Green & Nguyen, 2001; Utley et al., 2003; Bekker & Koeleman, 2011) and Markov Decision Process (MDP) models (Helm, AhmadBeygi, & Van Oyen, 2011; Barz & Rajaram, 2015; Samiedaluie, Kucukyazici, Verter, & Zhang, 2017; Li, Liu, Geng, & Xie, 2018), because they allow modeling the admission dynamics at the operational level. Still, such models are based on the assumption of a steady-state system and applied to a single hospital unit, due to the "curse of dimensionality"; a common drawback in such optimization models.

Finally, a small number of studies have applied, stochastic optimization (Min & Yih, 2010b), robust optimization (Mittal, Schulz, & Stiller, 2014), and distributionally robust optimization (Meng et al., 2015; Mak et al., 2014; Zhang et al., 2017), to improve hospital admission planning under bed capacity constraints. In several fields, it has been shown that such methodologies, when considered in a multi-stage fashion, guarantee consistency in decision-making at different planning levels. See, for instance, Zymler (2010), Shang and You (2018), Alvarez et al. (2020), for applied studies in risk management, manufacturing processes, and forest industry, respectively.

Overall, despite the clear dependence between the different temporal levels of decision, few contributions consider a hierarchical planning approach to perform admission plans. Most research focuses on solving problems at a single level of

decision, mostly the operational level (Hulshof et al., 2012). Thus, current literature lacks coordinated decision frameworks, which provide feedback from lower levels (subject to uncertainty) to higher levels to guarantee consistency of the global admission plan. In healthcare capacity planning, these inconsistencies not only generate extra management costs but also affect the service quality that translates into transfers, rejections, long waits, early death, among others inefficiencies.

Summarizing, admission planning in public hospitals is complex due to the uncertainty and limited resources. The uncertainty is mostly associated with natural processes that can not be eliminated but managed. Another factor is that the critical component to improve is the person's well-being, which directly impacts their quality of life. Also, the planning process is associated with decisions that impact public policies, so the improvements to be implemented are limited by the national healthcare budget expenses. While vast work has been done on bed capacity planning, the current state of the art in the scope of the admission planning problem requires more integrated and robust decision models to achieve the expected service level and resource utilization goals.

This dissertation focuses on the study of *intertemporal planning* for the admission planning problem, which is subject to uncertainty and constrained for resource capacity. We consider optimization methods under uncertainty in a multi-stage fashion to improve temporal consistency and coordination at the tactical-operational levels. We account for the inpatient service (i.e., secondary and tertiary levels of care) of public hospitals. The ward beds - defined as an approximation of hospital capacity - are the primary resource of planning due to its interaction with the overall flow of admission.

Some questions that arise are: (i) Which optimization methodologies under uncertainty can better handle decisions in admission planning while finding a balance between robustness and consistency? (ii) How to develop robust admission tactical plans

consistent with the operational level? (iii) What is the trade-off between the level of service offered to patients and resource utilization targets? (iv) How to improve the admission process under limited distributional information of uncertainties?.

1.5. State of the art

The admission planning problem contributions comprise studies in many disciplines, including operation research, statistics, public health, and medicine (Hall, 2012). Our focus is on studies related to techniques in the field of Operations Research and Management (OR/OM), which has been widely applied to healthcare services in recent years. OR/OM provides the necessary tools and methods to support decision-making in health care delivery, to improve the quality of services by considering limited resources.

The state of the art of the admission planning and control in the inpatient service is sub-divided into three main streams: (i) surgery planning, through the management of ORs; (ii) admission planning with bed capacity constraints; and (iii) surgery planning with downstream capacities, e.g., beds. The streams share several similarities and differences that we would like to highlight. The surgery planning stream focused on studies about OR allocation and scheduling at different levels of planning (Cardoen et al., 2010). Allocation decisions determine how many ORs to open and surgery-to-OR allocation, while scheduling decisions assign time intervals between surgeries (typically single OR). For the admission planning with bed capacity constraints, by contrast, allocation decisions aim to decide which patient group (i.e., classified by specialty) to admit. Scheduling decisions, in turn, address the problem of determining the time and date of admission for individual patients, and the assignment of patients to beds or rooms for their stay. In surgery planning, the ORs are resources that involve fixed and variable costs, so it is necessary to decide which ORs to open at a specific time before scheduling. On the

contrary, in admission planning, the beds are a fixed resource; thus, the decisions are within a different scope.

The surgery problem has also been studied with downstream bed capacities (Beliën & Demeulemeester, 2007; Beliën et al., 2009; Ceschia & Schaerf, 2016) aiming to level bed occupation. Through these studies, it has been shown that the integrated approach of master surgery scheduling with downstream bed capacities results in a better performance of the planning system since they are related operations. Nevertheless, most of these studies address bed allocation to a particular care service, usually ICU.

In this dissertation, however, we focus on studying a more general hospital system where the admission process is managed by a central unit of planning that receives admission requests from different care units. Therefore, it is expected to optimize the overall admission process and not a particular case of single care service. Thus, we are focused on the second stream, i.e., the admission planning problem in the inpatient service with bed capacities. Notwithstanding, we have included the other research lines in the literature review, since the mathematical modeling structure and scope are similar, although the objectives differ in many ways.

This section provides a state of the art of the relevant literature related to this thesis. Since there exists a vast distinction according to features and scope of decisions, we firstly account for studies related to applications at each hierarchical decision level, (i.e., strategic, tactical, operational) in Subsection 1.5.1. Subsequently, in Subsection 1.5.2, we describe studies considering integrated frameworks, i.e., papers that address the APP, including more than one decision level by using methodologies under uncertainty. The literature review is summarized in Table form at the end of each subsection, comparing the main characteristics of the admission planning problem with the up-to-date state of the

art. Finally, in Subsection 1.5.3, we discuss the main gaps and limitations identified in the literature review.

1.5.1. The Admission Planning Problem

Within the OR/OM field, the APP has been studied at different temporal levels of planning according to the scope of decisions. We employ the vertical axis of the taxonomy presented in Hulshof et al. (2012) to summarize the main contributions and optimization methodologies at each planning level for the inpatient service. Our classification is a refinement of the capacity planning stream that mainly considers beds as planning resources. Besides, we include some works that study ORs' management due to its interaction with admission decisions in the inpatient service and, therefore, similarities in the modeling approach.

At the *strategic level*, admission decisions account for bed capacity dimensioning of care services for long-term planning, based on resource utilization targets. Aggregated historical data of patient arrival and LoS is used to establish bed requirements. At this level, determining bed capacity is critical, given that it imposes constraints over the daily use of beds at the operational level. Sizing decisions are commonly based on average bed occupancy, historically defined as 85%. The bed occupancy target aims to measure hospital performance from which government entities decide the number of beds to be granted to a hospital or care unit. However, as stated in Green (2002), this indicator usually gives the wrong idea of excess capacity, causing the reduction of beds, and therefore, deficiency in the service offered to patients.

Simulation models have been applied to determine the size of care units or several departments in a hospital, aiming to improve the patient flow and prevent blockage (Harper & Shahani, 2002; Nguyen et al., 2005; Kokangul, 2008; Zhang et al., 2012). In Harper and Shahani (2002), several admission policies are studied to improve hospital efficiency

in terms of bed utilization and refusal rate. The authors emphasize the need to incorporate the necessary details (e.g., random arrival and patient stay) to calculate bed requirements in a hospital. For long-term capacity planning, Zhang et al. (2012) developed a simulation model to determine the number of beds needed over a multi-year planning horizon, to guarantee expected levels of waiting time. Overall, simulation techniques provide strong scenario analysis capabilities to imitate the complex dynamic of healthcare systems. Still, they are based on the assumption of full knowledge of the probability distributions and are hard to validate. An extensive overview of the use of discrete event simulation for capacity planning can be found in Günal and Pidd (2010) and Baru et al. (2015).

Queuing models also have been used to determine the size of a care unit. The main advantage is that such models allow modeling time-dependent stochastic flows when limited data is available. However, its primary assumption is based on the system's steady-state, which is not always applicable for finite planning horizons. The models assume random arrivals according to a Poisson process and exponential distributions of the patient LoS. For instance, Green and Nguyen (2001) proposed a queuing model to study how the use of utilization targets affects admission delays. The authors found that using fixed targets of resource utilization may lead to excessive delays for clinical units of small size. Utley et al. (2003) aiming to reduce the number of cancellations for acute patients, developed a queuing model to regulate the patient flow of acute and non-acute patients by creating an intermediate care unit and determining its size. The reader is referred to Lakshmi and Iyer (2013) for an extensive literature on queuing theory methods in bed capacity planning.

The admission planning at the *tactical level* has the objective of providing patient access (i.e., service concept) by controlling waiting and idle time and maximize resource utilization. It also determines the mix of patients admitted in a time horizon considering

resource utilization metrics and service-related metrics. Most of the studies have focused on resource utilization rather than the service concept in the planning process (J. M. Vissers et al., 2007). For instance, reference Adan et al. (2009) proposes a Mixed Integer Linear Program (MILP) model to determine the mix of patients to be admitted aiming to minimize the deviation for a targeted utilization considering different resources in the patient flow (e.g., beds, operating room, and personnel staff). The authors conclude that using stochastic values of the patient's length of stay will result in a better balance of resource use than the average length of stay. Similarly, Hulshof et al. (2013) studied an integrative approach to allocating resource capacities (i.e., OR) to care processes. The authors propose an iterative method based on MILP to determine the mix of patients to be served, considering the performance measures of resource utilization and balance workload.

Unlike focusing on improving resource utilization, the service concept is related to the management of waiting times and priorities rules to admit patients differing by predefined characteristics. To this matter, several schemes have been developed in the literature to define the priority levels by patient type (Mullen, 2003). Thus, due to the patient heterogeneity in the admission process, deciding which patient to admit is one of the most challenging issues. The decisions have welfare implications that must be taken into account. For instance, Testi and Tànfani (2009), and, Durán et al. (2017) evaluate the welfare implications of patient allocation for the OR problem. The authors employed a service measure to adjust the patient's waiting time to prioritize their allocation.

Two major patient categories are considered in the literature: elective and emergency patients. The planning of elective patients is more studied since it can be planned in contrast to emergency patients that require dynamic allocation rules or reserve policies of admission (Hulshof et al., 2016). Nevertheless, in practice, admission decisions concern

more than two categories of patients differing in their characteristics such as length of stay, level of illness severity, and other clinical issues. A qualitative study by Gemmel and Van Dierdonck (1999) states the importance of considering different patient types in the admission process to capture the inherent variability in the use of resources. A limited number of studies have developed the admission problem taking into account this broad of patient categories. Relying on this idea, J. Vissers et al. (2005) analyzes the APP considering ten patients types depending on the surgery duration and length of stay for a cardiothoracic service. Min and Yih (2010a), on the other hand, sub-classifies elective patients in urgent patient groups to achieve an optimal policy in operating room scheduling. The authors highlight the dependence of different patient priorities in the scheduling process.

Due to the complex interaction between scheduled and unscheduled patients, allocation policies must be flexible to ensure service levels; the technical literature study reservation policies to deal with this problem. For instance, Seung-Chul and Ira (2000) applied bed reservation policies based on a multi-objective simulation method, which measures different performance indicators simultaneously. The authors conclude that there is not a dominant solution to the problem of bed allocation. However, they offer a general solution scheme in which the trade-off is obtained between the impact on waiting time for canceled surgeries and the occupancy rate. In order to determine the optimal number of elective admissions, a queuing model is proposed in Bekker and Koeleman (2011), considering variability in patient arrivals and LoS. Their experiments showed that variability in admissions leads to higher variability in bed demand and more refused admissions. The authors propose an admission policy to establish a smooth admission pattern during the week to reduce bed utilization variation. Barz and Rajaram (2015) studied admission planning as a MDP considering multiple resource constraints and

uncertain care requirements. To address the curse of dimensionality, the authors propose an Approximate Dynamic Programming (ADP) based on heuristics.

At the *operational level*, the contributions focus on developing operational strategies on handling the patient's over-occupation due to limited capacity. The operational admission plan is carried-out weekly to daily and can be performed *online* or *offline*. The former is implemented while the system is ongoing and requires dynamic allocation of patients (see, e.g., X. Wang et al. (2018); Vancroonenburg et al. (2016)); the latter assumes the admission requests are known before the plan is made and considers non-emergency (elective) patients. Most contributions in the literature are focused on the offline models because elective patients represent a more significant amount of admissions and, therefore, more complexity in the decision-making. Within offline scheduling, studies concern the application of transfer strategies and patient's reassignment to other care units to manage unit congestion, usually applied to individual care units in a hospital.

Different sources of variability and uncertainty are present in the operational admission plan, such as patient arrival, capacity availability, and length of stay. Since uncertain parameters are difficult to estimate, most contributions consider estimates of the probability distributions of the uncertain parameters (He et al., 2019). In this regard, Kortbeek et al. (2015) proposed an analytic method to predict bed availability on an hourly basis for the problem of nurse staffing. The method is an interesting approach to be used as an input for operational models to improve the estimation precision of the uncertain parameter. In order to address variability, most studies apply analytical approaches such as queuing theory and MDP models (Samudra et al., 2016) because they allow modeling the admission dynamics at the operational level. Still, it can only be applied to an individual care unit or hospital department. For instance, Gallivan et al. (2002) proposed a queuing model to study the interactions between the LoS variability, booked admission,

and capacity requirements for an ICU. The authors conclude that a high degree of reserve capacity is required to cope with variability in the LoS.

To capture the dynamics of the admission process and manage the congestion in care units, Thompson et al. (2009) studied the relocation problem solved as a finite-horizon MDP. It was concluded that proactive patient relocation reduces waiting times, which contributes to both quality and efficiency. A similar analysis is found in Dobson et al. (2010) in which a discrete Markov chain approach is proposed. The objective is to evaluate the effect of transferring patients to other care units under different arrival patterns and capacity availability. A real-time allocation model was developed considering ICU beds. Different scenarios were analyzed to determine the effects of LoS variation on the patient's bumping decisions. A more detailed approach is studied in Hulshof et al. (2016) using an ADP framework. Many features are included that make the approach realistic to be applied in real-world situations. The authors concluded that ADP models are suitable for tactical admission planning, and can be used to readjust the plan by including new information related to arrivals and capacity availability.

Studies in dynamic strategies of bed reservations are found in Helm et al. (2011). The authors propose an analytic approach of MDP to smooth resource utilization by considering the beds as a shared resource of emergency and elective patients. The priority rules of admission are dynamic and related to the patient's waiting time in a queue. The authors use the results as an input to the strategic plan in a MILP model to obtain optimal admission schedules. A similar study is proposed in Liu et al. (2019) in which the authors also consider the utilization of inpatient beds for patient admission but simultaneously with OR utilization to study a two-stage hospital service system, considering ward transfers.

In general, studies applying dynamic approaches employing a MDP model are useful to capture the admission dynamics at the operational level, especially to determine whether

or not to accept patients belonging to a waiting queue. However, it is usually assumed full knowledge of the probability distribution of the uncertain parameters; Poisson arrivals and exponential distributions of the patient length of stay. See Gupta and Denton (2008) and He et al. (2019) for a detailed overview of the use of MDP for appointment scheduling in healthcare.

Due to the APP's combinatorial complexity, some authors have applied heuristic and metaheuristic-based methods as approximation algorithms. Although the models are computationally more efficient than integer programming methods, the solutions are only lower bounds on the optimal solution. Demeester et al. (2010) is among the first studies to introduce the APP as a combinatorial problem. The authors proposed a tabu-search algorithm that considers medical needs and patient preferences. The objective is to minimize the weighted sum of the penalty for assigning patients to no preferred rooms and the number of transfers. Demeester et al. (2010) found that the proposed algorithm outperforms previous Mixed Integer Programming (MIP) formulations, which takes over three hours to find a feasible solution. Nevertheless, the study assumes known admission dates and expected patient's length of stay. Ceschia and Schaerf (2012), by contrast, proposes a simulated annealing-based metaheuristic based on local search incorporating uncertain length of stay. The model, which is the same defined in Demeester et al. (2010), captures interesting features of real operations, such as room gender and age policy, specialty, and room equipment requirements. Intending to improve the solution time, Range et al. (2014) introduces an optimization-based heuristic of the APP based on a column and constraint generation approach. The author reported good performance in some instances.

A dynamic version of Ceschia and Schaerf (2012) was studied in Vancroonenburg et al. (2016). To model the daily dynamic of arrivals and departures of elective and

emergency patients, the authors consider estimates values of arrivals and departure dates, and patient LoS, employing an Integer Linear Programming (ILP) model. Experimental results indicate that including information about future arrivals in the admission plan yields better solutions in terms of admissions costs. The authors note as future research the study of scheduling and allocation simultaneously, instead of considering the admission date as an input parameter. Later, Turhan and Bilgen (2017) considered a MIP heuristic-based approach for solving the APP. The model was addressed for several instances while decomposing the patients based on their preferences and length of stay. The solutions of the approach are shown to be feasible in low computation times, but no optimality guarantee is obtained. Related studies can be found in, Ceschia and Schaerf (2016) and Guido et al. (2018). An exact solution method of a similar model was proposed in Bastos et al. (2019) as a MIP approach, achieving improved performance.

Overall, the studies mentioned above provide a significant contribution in terms of modeling and practical application. However, they lack integration between the temporal levels of planning, which would help to ensure consistency in decision-making. A summary of the contributions highlighting the main aspects relevant to this thesis is shown in Table 1.1. We group the studies by the level of planning, i.e., strategic, tactical, operational, in column 1. Columns 2 list the authors, while columns 3–11 show the main features of the APP. The number and hyphen symbols in each column indicate the presence or absence of a given characteristic defined at the bottom of the table. The features indicated in Table 1.1 are defined below.

(i) *Decision level:* is referred to as the level of decision in the APP, i.e., strategic, tactical, operational (offline-online).

(ii) *Type of problem* categorizes between scheduling and allocation decisions, which are sub-task in admission planning. We remark that allocation decisions aim

to decide which patient to admit considering bed availability, and scheduling decisions address the problem of determining time and date of admission and assigning patients to beds or rooms for the duration of their stay. For surgery planning, allocation decisions determine how many OR to open and the surgery-to-OR allocation, while scheduling decisions assign time intervals between surgeries (typically single OR).

(iii) *Resource type:* specifies which resource was considered for the admission problem, e.g., beds, ORs, or general services.

(iv) *Resource dimension:* indicates the use of single or multiple resources in parallel in the modeling framework.

(v) *Capacity availability:* classifies the contributions according to whether they consider time-varying or fixed bed capacity in the time horizon.

(vi) *Specialty dimension:* indicates the consideration of single or multiple care units or specialty in the modeling framework.

(vii) *Patient category:* indicates whether patients categories were considered, e.g., by grouping them according to their LoS or source of arrival or priority.

(viii) *Uncertain parameter:* shows the stochastic parameter considered in the modeling framework, e.g., length of stay, arrival.

(ix) *Solution method:* indicates the mathematical methodology used to solve the problem.

(x) *Criteria:* describes the solution criteria of the modeling framework.

TABLE 1.1. Summary of the literature on the APP: Comparison of main characteristics.

Decision level	Research	Type of problem	Resource type	Resource dimension	Capacity availability	Specialty dimension	Patient category	Uncertain parameter	Solution method	Criteria
Strategic (1)	Green and Nguyen (2001).	3	1	1	1	1	1	1-2	8	1
	Harper and Shahani (2002).	3	1	1	1	2	2	1-2	7	9
	Utley et al. (2003).	3	1	1	1	1	1	1-2	8	8
	Zhang et al. (2012).	3	1	1	1	1	1	1-2	7	1
Tactical (2)	Adan and Vissers (2002).	1	1-2	2	1	1	2	-	5	10
	Adan et al. (2009).	1	1-2	2	1	1	2	1	5	10
	Barz and Rajaram (2015).	1	1-2	1	1	1	2	2	9	7
	Bekker and Koeleman (2011).	1	1	1	1	1	2	1-2	8	10
	B. Denton and Gupta (2003).	2	2	1	1	1	1	4	6	1-3
	Hulshof et al. (2013).	1	3	2	1	2	2	2	9	1
	Hulshof et al. (2016).	1	3	2	1	2	2	2	9	1
	Seung-Chul and Ira (2000).	1	1-2	2	1	1	1	2-3	7	1-9
Offline (3)	Bachouch et al. (2012).	2	1	2	1	1	2	-	5	9
	Ceschia and Schaerf (2011).	2	1	2	1	2	1	-	4-9	8
	Conforti et al. (2011).	2	1-2	2	1	1	2	-	5	7
	Demeester et al. (2010).	2	1	2	1	2	2	-	3	8
	Guido et al. (2018).	2	1	2	1	2	1	-	3-5	8
	Range et al. (2014).	2	1	2	1	2	1	-	9-10	8
	Turhan and Bilgen (2017).	2	1	2	1	2	2	-	4	8
Online (3)	Ceschia and Schaerf (2012).	2	1	2	1	2	2	1	3-5	8
	Ceschia and Schaerf (2016).	2	1-2	2	1	2	2	1-2	3-4	8
	Helm et al. (2011).	2	1-2	1	1	1	2	2	5-11	10
	Li et al. (2018).	2	1	2	1	1	2	1-2	11	11
	Liu et al. (2019).	1	3	1	1	1	1	1-2	11	10
	Mazier et al. (2010).	2	1	2	1	1	1	2	5-7	1-8
	Samiedaluie et al. (2017).	2	2	2	1	1	2	1-2	9	1
	Vancroonenburg et al. (2016).	2	2	2	1	2	2	1	3-5	8

Decision level: 1 (strategic); 2 (tactical); 3 (operational). - **Type of problem:** 1 (allocation); 2 (scheduling); 3(sizing) - **Resource type:** 1 (ward beds); 2 (OR); 3 (general jobs). - **Resource dimension:** 1 (single); 2 (multiple). - **Capacity availability:** 1 (constant); 2 (variable). - **Specialty dimension:** 1 (single); 2 (multiple). - **Patient category:** 1 (single); 2 (multiple). - **Uncertain parameter:** 1 (length of stay); 2 (demand/arrival); 3 (service time); 4 (surgery duration); 5 (No - shows). - **Solution method:** 1 (DRO); 2 (Chance constraint); 3 (meta-heuristics); 4 (other heuristics); 5 (ILP); 6 (stochastic programming); 7 (simulation); 8 (queuing theory); 9 (dynamic programming); 10 (BIP); 11 (MDP); 12 (robust optimization). - **Criteria:** 1 (waiting time); 2 (overtime); 3 (idle time); 4 (opening OR); 5 (bed shortages); 6 (overstay); 7 (admission benefit); 8 (admission cost); 9 (cancellation); 10 (resource utilization); 11 (others).

1.5.2. Intertemporal planning and decision-making under uncertainty

Limited research has been developed that considers intertemporal approaches for the admission planning problem with bed capacity constraints. Some studies have instead

studied two-steps frameworks to solve the APP considering different levels of planning. For instance, Adan et al. (2011) proposed a MIP model divided into two sequential stages; the first stage decisions reserve beds for urgent patients while the second stage decisions are operational strategies to manage the flow of elective and urgent patients. The results show the relationship between patient satisfaction and resource utilization. Nevertheless, the results are obtained through simulation, applying flexibility rules between elective and urgent patients. Thus, there is no guarantee of consistency in intertemporal decision-making.

One way to ensure consistency between different planning levels is through the application of decision-making methods under uncertainty in a multi-stage fashion. Such methods guarantee that the aggregate plan at upper levels can be efficiently implemented at lower levels at a certain cost.

The most common method to handle uncertainty in an intertemporal fashion is two-stage stochastic programming, which assumes full information about the distribution of uncertain parameters (Birge & Louveaux, 2011). For the APP in the inpatient service, few contributions are found in the literature; studies are mainly applications of surgery care services, considering in some cases, downstream bed capacities under uncertain surgery duration. For instance, in the context of surgery planning, B. Denton and Gupta (2003) proposed a TSO approach for the appointment scheduling problem considering a single OR. They aimed to determine the planned start times while minimizing OR waiting and idle times, under stochastic surgery durations. The authors considered the Sample Average Approximation (SAA) and L-shaped algorithms to solve the model. An interesting finding is that the job allowances tend to take a dome-shaped structure when idling costs are high in comparison to low waiting costs. On the contrary, a uniform distribution is found for the opposite case of high waiting cost and low idling costs.

Later, B. T. Denton et al. (2010) extends the model presented in B. Denton and Gupta (2003) by considering multiple ORs instead of single resource. Min and Yih (2010b) also proposed a TSO model for elective surgery scheduling but under downstream bed capacities. It assumes uncertainty in surgery durations and LoS in ICU beds. The authors reported optimal solutions when employing a SAA with a moderate sample size. As a future research direction, it is proposed the study of different allocation cost structures considering the patient medical condition. Similar approaches are found in Jebali and Diabat (2015, 2017).

More recently, Vancroonenburg et al. (2019) considers downstream capacities for patients to stay after surgery, including the daily dynamic patient allocation. The authors propose a stochastic chance-constrained model to minimize OR allocation costs and patient access for a single specialty. It is particularly interesting that their model accounts for real-time decisions, in which the schedule is revised in response to future changes. Such a dynamic setting is modeled through a heuristic local-search algorithm. Analogous to the contributions mentioned above, Vancroonenburg et al. (2019) addresses the surgery scheduling problem employing approximated methods and heuristics, assuming full knowledge of the uncertain parameters. The authors point out that under a constrained allocation process under uncertainty, a balanced performance can be obtained at the expense of higher waiting times and late notification to patients.

In general, the TSO approach is reported to be effective in terms of performance to solve the APP, albeit neutral risk. Of course, this method's effectiveness is subject to the accurate definition of the uncertain parameters, which should be properly estimated to avoid adverse results due to inaccurate estimations.

Robust optimization is an alternative framework in the absence of information about the true distribution of the uncertainties. In RO, uncertain parameters are described by

means of uncertainty sets, making no assumptions about the probability distributions (Ben-Tal & Nemirovski, 1998; Bertsimas & Sim, 2004). In particular, adaptive robust optimization is solved in several stages, aiming to find a robust plan under the worst-case scenario within the predefined uncertainty set, (Bertsimas, Litvinov, Sun, Zhao, & Zheng, 2012). For the APP, very few contributions have been developed under this framework. We only found the contribution of Mittal et al. (2014), who studied the appointment scheduling problem to minimize waiting and idle time costs for general jobs under uncertain processing times. The authors assume partial information for the service durations that lie in an interval uncertainty set. A heuristic is developed to balance the costs of jobs in the allocation process. Although the authors proved to obtain optimal solutions under the RO approach, such models sometimes could lead to very over-conservative solutions limiting its utility in the healthcare setting.

Recently, distributionally robust optimization has received increasing attention in many fields (Bertsimas et al., 2011; Gabrel, Murat, & Thiele, 2014). DRO models allow more information to be included in the uncertainty set, such as the mean and covariance, that can be estimated from historical data or expert knowledge (Wiesemann et al., 2014). For the healthcare system, the DRO scheme of decisions is relevant since it is expected to prioritize the patient's welfare under robust decisions, but not over-conservative, as in the case of RO. Nevertheless, not many studies have considered the DRO approach in APP under bed capacity constraints.

The work presented in Mak et al. (2014) is among the first to study the APP with limited distributional information of service duration. For a single resource and a fixed sequence of appointment arrivals, the authors study the surgery appointment scheduling problem to minimize the cost of patient waiting times and overtime considering uncertain service time. The model is solved by deriving tractable conic reformulations assuming

a marginal moment (i.e., mean-support, mean-variance) ambiguity set of the random service time. The optimality of the order of variance policy, for patient sequencing, was proved analytically. The authors indicated, as further research, the problem of scheduling in conjunction with allocation decisions. Later, Jiang et al. (2017) solved the appointment scheduling problem presented in Mak et al. (2014) by incorporating ambiguity in heterogeneous no-shows in addition to service times.

A DRO MILP is proposed in Meng et al. (2015) by deriving a second-order conic programming counterpart, under moment ambiguity set. The objective is to determine (at the tactical level), quotas of admission for elective and emergency patients under patient arrival uncertainty. The DRO minimizes the worst-case expected bed excess in the planning horizon. Through simulation studies with real data and a rolling horizon approach, they concluded that the poor choice of the budget of uncertainty could lead to inferior solutions in a robust optimization approach. Additionally, they claim that bed shortfalls are inevitable in any hospital that is operating near its capacity.

The most recent study is found in Y. Wang et al. (2019), in the context of OR planning. The authors studied the surgery block allocation problem presented in B. T. Denton et al. (2010) as a DRO approach considering uncertainty in the surgery durations. The model aims to minimize the total cost of opening ORs assuming the overtime cost over the worst-case probability distribution within a moment ambiguity set. The authors reveal the importance of considering heterogeneous service time in the scheduling process to improve prediction accuracy.

Distributionally Robust Chance Constrained (DRCC) approaches are also considered to solve the admission planning problem under distributional ambiguity of the uncertain parameters. DRCC under ambiguity guarantees that the probability of meeting a certain constraint within the worst-case distribution is not above a risk tolerance (Zhang et al.,

2018). The approach has the advantage that it does not rely on estimations of cost parameters in its formulation, to reflect the impact of decisions at lower levels of planning.

Zhang et al. (2017) developed a DRCC model to solve the appointment scheduling problem. The authors assume ambiguity of service times distribution, which follows a moment-based approach. The objective is to minimize the cost of waiting time by restricting the risk of overtime and considering a fixed order of arrivals. In particular, the authors incorporate ambiguity in both the objective function (i.e., waiting time) and the chance constraint of overtime. The model is solved by reformulating the chance-constrained formulation as an equivalent semidefinite programming (SDP) model through dual theory. When compared to a sampling-based stochastic linear approach, they found that a DRCC model ensures lower levels of overtime with a high probability.

Later, Zhang et al. (2018) proposed a similar model for surgery block allocation, aiming to minimize the total cost of opening ORs due to overtime. In contrast to the approach presented in Zhang et al. (2017), the reformulation to an SDP model includes the binary variables of opening OR, thus, in order to handle the 0–1 SDP reformulation the authors proposed a cutting-plane algorithm and a 0–1 second-order cone program (SOCP) approximation. The authors conclude that the 0–1 SDP approach yielded better solutions than the SOCP approximation.

The authors in Deng et al. (2019), in contrast to previous contributions, solved an integrated model of scheduling and surgery-to-OR allocation to determine the sequence of performing surgeries and decisions of opening ORs. A DRCC model is proposed to constrain overtime and waiting times, under a phi-divergence ambiguity set. They define the decisions of opening ORs and start-times as binary variables and the sequence of allocation, as continuous. A SAA approach is employed to reformulate the chance constraint problem as a MILP solved by a branch-and-cut algorithm. The authors

provide insights into the relevance of developing integrated frameworks of scheduling and allocation to get a better balance in the use of resources in operational planning. We refer to Delage and Ye (2010) for further details about applications in DRCC under a moment ambiguity set.

Overall, most contributions applying intertemporal approaches considering methods under uncertainty, have been studied in the context of surgery planning. More research is needed to derive robust plans of admission under bed capacity constraints. Similar to the previous subsection, we summarize in Table 1.2 the main contributions that have considered integrated or hierarchical planning approaches to solve the admission planning problem.

TABLE 1.2. Summary of literature on the APP under intertemporal planning: Comparison of main characteristics.

Research	Decision Level	Type of problem	Resource type	Resource dimension	Capacity availability	Specialty dimension	Patient category	Uncertain parameter	Solution method	Ambiguity set	Criteria
Deng et al. (2019).	2-3	1-2	2	2	1	1	2	4	1-2	4	1-2
B. T. Denton et al. (2010).	2-3	1	2	2	1	1	1	4	6	-	2
B. Denton and Gupta (2003).	2-3	2	2	1	1	1	1	4	6	-	1-3
Jiang et al. (2017).	2-3	2	3	1	1	1	1	4-5	1	1-2	1-2-3
Mak et al. (2014).	2-3	2	3	1	1	1	1	3	1	1-2-3	1-2
Meng et al. (2015).	1-2	1	1	1	1	1	1	2	1	1-2	5
Min and Yih (2010b).	2-3	1	1-2	1	1	1	1	1-4	6	-	2-8
Mittal et al. (2014).	2-3	1	3	1	1	1	1	3	12	-	1-3
Vancroonenburg et al. (2019).	2-3	1-2	1-2	2	1	1	2	1-4	2-4	-	8
X. Wang et al. (2018).	2-3	1	2	2	1	1	1	4	1	1-2	2-8
Zhang et al. (2017).	2-3	2	3	1	1	1	1	3	1-2	1-2-3	1-2
Zhang et al. (2018).	2-3	1	2	2	1	1	1	4	1-2	2-3	2

Decision level: 1 (strategic); 2 (tactical); 3 (operational). - **Type of problem:** 1 (allocation); 2 (scheduling); 3(sizing) - **Resource type:** 1 (ward beds); 2 (OR); 3 (general jobs). - **Resource dimension:** 1 (single); 2 (multiple). - **Capacity availability:** 1 (constant); 2 (variable). - **Specialty dimension:** 1 (single); 2 (multiple). - **Patient category:** 1 (single); 2 (multiple). - **Uncertain parameter:** 1 (length of stay); 2 (demand/arrival); 3 (service time); 4 (surgery duration); 5 (No - shows). - **Solution method:** 1 (DRO); 2 (chance constraint); 3 (meta-heuristics); 4 (other heuristics); 5 (ILP); 6 (stochastic programming); 7 (simulation); 8 (queuing theory); 9 (dynamic programming); 10 (BIP); 11 (MDP); 12 (robust optimization) - **Criteria:** 1 (waiting time); 2 (overtime); 3 (idle time); 4 (opening OR); 5 (bed shortages); 6 (overstay); 7 (admission benefit); 8 (admission cost); 9 (cancellation); 10 (resource utilization); 11 (others).

Column 1 lists the authors, while columns 2–12 show the relevant features of the study. In comparison with Table 1.1, we have added column 10 to identify the

feature, *Ambiguity set*, which characterizes the studies that consider a DRO approach as a resolution method. The number and hyphen symbols in each column indicate the presence or absence of a given characteristic defined at the bottom of the table.

1.5.3. Literature gaps and limitations

From the literature review presented in Subsections 1.5.1 and 1.5.2, we have identified relevant limitations in the study of the admission planning problem under bed capacity constraints. In particular, three main gaps are evidenced, namely (i) the lack of an integrated framework to guarantee temporal consistency in the admission planning, (ii) the study of multi-objective criteria to evaluate the trade-off between conflicting objectives typically found in the admission planning problem, and (iii) the study of intertemporal modeling approaches through optimization methods under uncertain length of stay to guarantee robust plans of admission. Those aspects will be described in detail in the following subsections.

1.5.3.1. Use of an integrated framework in the Admission Planning Problem

The APP problem has been approached in many ways at different levels of temporal planning. The existing literature is focused on optimizing patient throughput, considering criteria of waiting time, resource utilization, and total admission cost. As we showed in Table 1.1, most papers focus on a single level of decision. Thus, no consistency guarantee or coordination between decision levels is acknowledged. From the studies, Adan et al. (2011) considers the hierarchy in the decision-making for the surgery planning problem. However, the approach is a two-step model that is solved sequentially, not in an integrative framework, which may lead to infeasibility in the short-term.

In general, most contributions focus on one decision level without considering the impact of future realizations when making decisions at higher levels. Thus, the applied

admission policies would result in inconsistencies during the plan execution, such as rejections and unnecessary waiting times.

1.5.3.2. Multi-objective criteria to evaluate conflicting objectives

The decisions in the APP under bed capacity constraints are driven by several conflicting objectives, divided into two broad categories: the *hospital and patient perspective*. The hospital aims to obtain the maximum performance of bed utilization. At the same time, the patient expects to receive the best service quality, measured as a reduction in delays, cancellations, and waiting times. The study of both objectives as part of an intertemporal multi-objective approach, has not been studied. To date, only one contribution, Seung-Chul and Ira (2000), have considered a multi-objective framework, evaluating a Pareto-frontier to trade-off canceled surgeries and waiting time, as a result of bed reservations. However, they performed the analysis via simulation, for one specialty, and at a single level of planning (operational), which may lead to infeasible solutions.

1.5.3.3. Intertemporal approach through optimization methods under uncertainty

Most existing optimization models employing an intertemporal approach for admission planning, have been applied in the context of surgery planning. We remark that these papers do not specify the use of hierarchical models as the main contribution, but the modeling structure implies it. The studies consider, the OR (Deng et al., 2019; B. T. Denton et al., 2010; Y. Wang et al., 2019; Zhang et al., 2018) or both OR and beds as downstream capacity (Min & Yih, 2010a; Vancroonenburg et al., 2019). Such works are focused on developing robust models to overcome the consequences of surgery duration uncertainty. Other papers do not explicitly indicate the resource being planned[1] (Jiang et al., 2017; Mak et al., 2014; Mittal et al., 2014; Zhang et al., 2017), and consider the service time as uncertain parameter, to reduce waiting and overtime in the allocation and

[1]Categorized as general jobs in Tables 1.1 and 1.2.

scheduling processes. Only one contribution, (Meng et al., 2015), considers ward beds as the main focus of planning, for the strategic-tactical level of decisions. In Meng et al. (2015), the patient's demand is assumed uncertain, aiming to minimize bed shortages and smooth bed utilization.

From table 1.2 we observe that while some studies assume perfect information of the probability distribution of the uncertain parameter, employing TSO (B. T. Denton et al., 2010; Min & Yih, 2010b) and chance constraint methods (Vancroonenburg et al., 2019), others assume ambiguity of the distribution using a DRO approach (Deng et al., 2019; Jiang et al., 2017; Mak et al., 2014; Meng et al., 2015; Y. Wang et al., 2019; Zhang et al., 2017, 2018), or RO (Mittal et al., 2014). Although the studies mentioned above model the admission planning problem, employing approaches that guarantee coordination between different decision levels, most partially cover the essential characteristics of the problem.

For instance, both of the contributions that consider a TSO approach (B. T. Denton et al., 2010; Min & Yih, 2010b) are focused on optimizing OR utilization, assuming a single patient category and resource to allocate time intervals between appointments. As these studies focus on surgery planning, aspects such as the deviation in the use of beds, multi-priority, and multi-specialty are not acknowledged. On the other hand, the contributions that applied a DRO approach, are mostly focused on allocation decisions, i.e., determining OR opening for patients to be admitted to a single OR (X. Wang et al., 2018; Zhang et al., 2018) or general service (Jiang et al., 2017). Still, not a multi-period, multi-priority, and multi-specialty approach is considered. Other studies, see, e.g., Mak et al. (2014) and Zhang et al. (2017), consider only scheduling decisions, determining time intervals between general jobs under stochastic service time. Most contributions, ignore the integrated framework of allocation and scheduling in the admission planning.

Relevant exceptions are the papers Deng et al. (2019), and Vancroonenburg et al. (2019) for a multi-priority scheme. However, they are studied in the context of surgery planning.

In summary, to date, few contributions in the technical literature study the admission planning problem under bed capacity constraints, to achieve consistency between temporal levels of planning by employing optimization methods under uncertainty and capturing real-life characteristics of the problem. There is also a need for integrated approaches to allocation and scheduling in the context of bed capacity planning; it has been shown that better performance can be achieved when both processes are integrated (Deng et al., 2019).

1.6. Thesis hypothesis and objectives

1.6.1. Hypothesis

The main hypothesis of the thesis is that it is possible to model and solve hierarchical models in multiple stages to improve decision-making in admission planning for the healthcare system.

The specific hypotheses are described as follows:

(i) Hierarchical decision frameworks allow obtaining robust plans that guarantee consistency between different decision levels in healthcare admission planning.

(ii) It is possible to develop mathematical optimization models that consider intertemporal approaches for admission planning in the healthcare sector to ensure consistency between decision horizons.

(iii) It is possible to find a trade-off between service level and resource utilization under uncertain conditions, to ensure flexibility in the decision-making that benefits conflicting parties.

(iv) It is possible to obtain robust solutions at the tactical level, which guarantees a balance between robustness and the consistency of decisions at the operational level, considering limited information of uncertain parameters.

(v) It is possible to provide empirical evidence on real data to enhance hospital admission operations through cost-efficient practical guidelines.

1.6.2. Objectives

The main objective of the thesis is to model and solve a hierarchical decision-making process, in multiple stages, for admission planning in the healthcare system, to improve temporal consistency under uncertain conditions.

The specific objectives are described as follows:

(i) To study hierarchical decision methodologies under uncertainty that allow to obtain robust admission plans considering bed capacity constraints in the healthcare sector.

(ii) To model and solve the problem of admission planning through an intertemporal approach that adjusts the needs of supply to demand and ensures consistency at the tactical-operational levels.

(iii) To develop an intertemporal mathematical model under uncertainty, which allows finding a trade-off between the level of service and the use of resources in admission planning.

(iv) To study the robustness of decisions in a hierarchical decision setting while finding a balance between robustness and consistency under limited information of the uncertain parameters.

(v) To provide empirical evidence on real data that support hospital admission operations through cost-efficient practical guidelines.

1.7. Structure of the thesis and contributions

1.7.1. Structure of the thesis

The remaining of this document is organized into six additional chapters. Each chapter aims to answer one or various objectives, as previously described in Subsection 1.6. We remark that the content of the Chapters 3–5 are based on individual papers published or submitted in journals. The mathematical notation is presented independently in each chapter. Chapter 6 concludes the thesis and gives future research directions. The document finishes with the Appendix.

The chapters content are outlined as follows:

Chapter 2 provides a summary of the background theory of decision-making optimization approaches under uncertainty that are used in this thesis. In particular, we describe the basic concepts of two-stage stochastic optimization, robust optimization, and distributionally robust optimization. Besides, we provide a brief up-to-date state of the art of decision-making methods under uncertainty and related solution methodologies.

Chapter 3 studies the allocation decisions in the admission planning problem, considering an intertemporal approach at the tactical-operational levels. A novel framework is presented as a TSO model in a multi-objective fashion. The approach allows evaluating the APP from both perspectives, hierarchical structure, and uncertain nature of the problem. The model defined as a mixed-integer linear programming problem includes reserving bed capacity decisions at the tactical level and patient allocation decisions at the operational level, constrained for demand and capacity availability uncertainties. The bi-objective approach, defined through the weighted-sum method, evaluates the trade-off between resource utilization deviation and cost of service. Real data from a Chilean public hospital illustrates the approach and validate the model. The results show that the solutions

of the proposed approach outperform the actual practice in the Chilean hospital. We provide insights to practitioners on balancing conflicting decisions of resource utilization and service level.

The content of this chapter is based on the paper, Batista, A., Vera, J., & Pozo, D. (2020), Multi-objective admission planning problem: A two-stage stochastic approach. *Health Care Management Science*, 23, 51-65.

Chapter 4 proposes an alternative framework for modeling service time constraints for problems that cannot be interrupted once allocated. Motivated by the admission planning problem with uncertain length of stay, this chapter provides a formulation framework that includes a single binary variable of service allocation and the service time on the right-hand side of the allocation constraint in a multi-period system. We contrast the proposed formulation with the current approaches in the literature. Besides, we describe the main variables and constraints of the proposed MILP formulation, which can be generalized for the type of problems of uninterruptible services in scheduling theory. The main advantage of the formulation is that the uncertain parameter is on the right-hand side of the constraint, rather than over the indexes of a summation as it is usually modeled. This feature facilitates the implementation of existing algorithms (e.g., dual-based methods, or Benders decomposition) that consider uncertainty, such as stochastic programming, robust optimization, and distributionally robust optimization.

The content of this chapter is partially based on the paper Batista, A., Pozo, D., & Vera, J. (2020), Stochastic time-of-use-type constraints for uninterruptible services. *IEEE Transactions on Smart Grid*, 11(1), 229-232.

Chapter 5 studies the robustness of decisions in an intertemporal decision framework for the admission planning problem. The modeling framework developed in Chapter 4 is employed to model a patient-to-room admission problem. The model aims to maximize

patient access and the use of available resources. In contrast to Chapter 3, in which the demand is considered uncertain, in this chapter, the demand is known, and the patient length of stay is uncertain. Besides, this chapter incorporates scheduling decisions in addition to the allocation decisions studied in Chapter 3.

Due to the lack of reliable information in most hospitals, we assume limited distributional information of the patient LoS. In order to solve the problem, a DRO framework is presented that is distribution-free; it considers that known information is limited only to the first moment and the support set of the true probability distribution. The framework is robust against the infinite set of probability distribution functions that could represent the stochastic process of the patient's length of stay. The resulting infinite-dimensional linear optimization problem is reformulated as an exact finite-deterministic mixed-integer linear problem. To demonstrate the effectiveness of the proposed approach, we compared it with benchmark models (i.e., deterministic, TSO, RO) employing a real data set from a public hospital in Chile. The results show that the DRO approach outperforms the benchmark models in both reliability and computational efficiency. It provides the most cost-efficient combination of consistency and robustness. We provide insights to practitioners and hospital decision-makers to anticipate admission decisions at the tactical-operational level while considering the randomness of the length of stay.

The content of this chapter is based on the paper Batista, A., Pozo, D., & Vera, J. (2020). Managing the unknown: A distributionally robust model for the admission planning problem under uncertain length of stay. *Computers and Industrial Engineering*.

Chapter 6 provides a summary, the conclusions and future research directions of the thesis.

Appendix A describes a data collection sheet designed to collect data about bed capacity requirements and availability of the hospital under study.

Finally, the **Appendix B** includes the data input employed in Chapter 5.

1.7.2. Contributions

1.7.2.1. General contributions

The general contributions of the research conducted in this thesis are listed below:

(i) Regarding admission planning considering a stochastic hierarchical framework. The thesis acknowledges the importance of considering the hierarchical and stochastic characteristics of the admission planning problem in the healthcare setting. Since most studies in the literature of healthcare capacity planning are developed at a single level of decision, the thesis fills this gap; the consistency issue between different decision horizons of the admission planning problem is studied. The thesis shows the advantages of an intertemporal decision framework to manage coordination between temporal levels of planning and to achieve robust solutions while considering variability and uncertainty at the operational level. By considering optimization methods under uncertainty in a multi-stage fashion, we improve consistency and coordination between tactical-operational planning.

(ii) Regarding admission planning considering alternative modeling frameworks. The thesis provides novel modeling frameworks for the admission planning problem under bed capacity constraints.

Firstly, the thesis considers a bi-objective TSO approach within a hierarchical framework of decisions. Besides considering the consistency issue between tactical and operational levels of decision, the modeling structure allows evaluating conflicting objectives commonly presented in the healthcare setting. Secondly, the thesis provides an alternative modeling framework that incorporates service-time-type constraints for services that cannot be

interrupted once allocated. The framework, which relies on a single binary variable, enhances current service allocation models under uncertain service duration. In particular, the proposed modeling is used to formulate the admission planning problem under uncertain patients length of stay. The thesis recognizes the lack of reliable information on the patient stay duration and presents an adaptive DRO-based formulation. The approach allows decision-makers to take admission decisions acknowledging the ambiguity of the true probability distribution of the uncertain parameter. Through this approach, the thesis also gives insights over the balance between robustness and consistency in a capacity-constrained system.

(iii) Regarding admission planning modeling considering data and insights from real hospital practice.

The thesis devises a practical value of the proposed mathematical optimization models and their applicability in the healthcare system. Contrary to most works in the literature that focus on theoretical approaches, the proposed modeling frameworks incorporate real-life characteristics that enhance their applicability in the healthcare system. The decision framework's effectiveness is demonstrated through extensive numerical studies and validation employing real data from a public hospital in Chile. From a managerial point of view, the thesis provides insights to practitioners and decision-makers to anticipate decisions at the tactical level while considering the randomness at the operational level. We detail several guidelines to derive decisions to favor patient and hospital perspectives instead of a single benefit.

1.7.2.2. Specific contributions

The specific contributions of each chapter of this thesis are listed below:

Chapter 2

(i) *Study hierarchical decision methodologies under uncertainty*

We study decision-making methodologies that allow modeling intertemporal problems under uncertainty. The chapter refers to the relevant literature in the field of operations research, which can be used as a guide to understanding the theoretical foundations in decision making under uncertainty.

Chapter 3

(i) *Develop an intertemporal stochastic approach for the APP.*

We propose a TSO model to address the APP for optimal patient allocation on beds at the tactical and operational levels, considering demand and capacity uncertainties. The approach allows evaluating the APP from both perspectives, hierarchical structure, and uncertain nature of the problem, to guarantee consistency in the admission process.

(ii) *Study the APP as a two-stage stochastic multi-objective problem.*

We incorporate a bi-objective approach to evaluating the trade-off between two conflicting objectives in the APP: resource utilization deviation and the cost of service. To the best of our knowledge, this study is the first effort in the literature to explore the APP as a two-stage stochastic model in multi-objective fashion. The model accounts for a balance of service level considering hospital and patient perspectives in the allocation process. We include flexible options for allocation, such as diverting patients to another hospital and temporary assignments. Also, unlike most studies, we consider bed allocation decisions for the entire hospital instead of a single unit.

(iii) *Validation of the proposed APP with real data.*

The proposed approach is validated with real practices on the APP for a

Chilean public hospital. Also, we provide insights to practitioners on balancing conflicting decisions of resource utilization and service level.

Chapter 4

(i) *Develop an alternative formulation for modeling service-time-type constraints of uninterruptible services.*

We propose a new but simple, effective formulation that includes service time constraints for problems in which the service's interruption is not allowed.

(ii) *Enhancement of service allocation models under uncertain service duration.*

We enhance current admission planning (or appointment scheduling) models by considering a single binary variable and continuous service time on the right-hand side of the formulation, rather than over the indexes of a summation. This structure facilitates the implementation of the existing algorithms (e.g., dual-based methods, or Benders decomposition) that consider uncertainty, such as stochastic programming, robust optimization, and distributionally robust optimization.

Chapter 5

(i) *Develop a new version of the APP that considers various patient types, multi-specialty, and time-varying bed capacity.*

We formulate a different version of the APP under stochastic patient length of stay. The model focuses on maximizing patient access to guarantee the use of available bed resources, although minimizing the cost of overstay. The approach is relevant for public hospitals where the bed costs are fixed (i.e., the resources are given a priori), and the main objective is to treat the most significant number of patients.

The model is solved at the tactical-operational level and provides insights into the relevance of considering scheduling and allocation decisions simultaneously; it determines the planned start times, time allowances, and bed assignment of patients. Unlike most studies in the patient-to-room admission problem, we do not assume a fixed sequence of arrival. Additionally, we capture the characteristics of the real-life setting in the modeling approach. For instance, we consider the heterogeneity of the patient LoS, including several patient types (i.e., multi-specialty, multi-priority) and room assignment with multiple identical parallel beds instead of a single resource. In contrast to previous research in which capacity availability is considered constant in the time horizon, we assume time-varying bed capacity (as it happens in practice), intending to model the real dynamics in the admission process.

(ii) *Extension of the admission planning model for considering LoS uncertainty for deriving a robust admission planning.*

We propose a distributionally robust optimization approach to solve the admission planning problem. Rather than assuming perfect information on the probability distribution of the patient LoS, we account for an ambiguity-averse framework. We consider that known information is limited only to the first moment and the support set of the true probability distribution of the LoS. This framework is convenient for the healthcare sector, lacking reliable information about the probability distributions of the uncertain length of stay and seeking to focus on high-quality care at a lower cost. Additionally, we derive a tractable solution methodology for solving the DRO through dual theory. Thus, the infinite-dimensional DRO is then reformulated into a deterministic equivalent model.

(iii) *Validation and thoughtful analysis of the proposed DRO admission planning model with real data. Comparison with standard TSO and RO frameworks for decision-making under uncertainty.*

To the best of our knowledge, we are among the very few contributions to account for real data under the proposed framework. The data, which is obtained from the Electronic Health Records (EHR) of a public hospital in Chile, is employed to construct the ambiguity set and generate the sample scenarios. We illustrate the robustness of the proposed approach through an extensive computational study by comparing it with standard approaches in the literature: robust optimization, two-stage stochastic programming, and deterministic. Besides, we propose a reliability metric by benchmarking with different approaches and conventional cost-based metrics in out-of-sample analysis.

Chapter 2. DECISION MAKING UNDER UNCERTAINTY

The decision-making process in many real-world problems is subject to uncertainty. This uncertainty can be due to several reasons, including predictable and unpredictable events such as future demand or the cost of a service or product. Problems subject to uncertainty can be formulated as optimization models that require the characterization of the unknown parameters. In some cases, the uncertainty is difficult to estimate, either because there is no adequate information or due to the fact that parameters that represent such uncertainty are very volatile. In the healthcare system, for instance, uncertain parameters are commonly associated with patient health and are difficult to estimate, such as the demand for unscheduled patients and the length of stay that, in turn, affects the availability of resources. Thus, failure to consider uncertainty in the planning process could result in suboptimal decisions (Zymler, 2010).

According to how uncertainty is taken into account, several optimization frameworks can be considered for capacity planning problems. Among these, we have deterministic methods, stochastic programming, robust optimization, and distributionally robust optimization. Deterministic approaches assume that uncertain parameters are known and can be estimated using expert knowledge or historical data. However, this approach could over or underestimate the solutions since decisions do not consider future changes caused by the realization of uncertainty. As for stochastic programming, the main assumption is that the probability distribution of uncertain parameters is known. A particular case is two-stage stochastic optimization, in which decisions are modeled in two stages. The expected value of the objective function is optimized on the scenarios that describe the uncertain parameters (Birge & Louveaux, 2011). The main drawback of this approach is that a significant number of scenarios are needed to characterize the uncertainty, so the solution could be computationally expensive. Unlike stochastic programming, robust

optimization requires minimal probabilistic information to estimate unknown parameters. Uncertainty is modeled as parameters that belong to a set of uncertainty (Bertsimas & Sim, 2004). One of the main disadvantages of robust optimization is that the solutions can be very conservative since they are driven by the worst-case within the set that represents the uncertainty. Finally, distributionally robust optimization is an alternative approach that takes into account distributional information of uncertain parameters. Thus, in addition to including the uncertainty set, it considers partial distributional information, such as the mean and variance. The solution, therefore, acknowledges the worst-case expected total cost instead of the worst-case scenario (Delage & Ye, 2010).

In this chapter, we summarize the background theory of decision-making under uncertainty to be employed in the subsequent chapters. In particular, we give a general description and overview of stochastic programming, robust optimization, and distributionally robust optimization methodologies. We remark that the concepts presented in this chapter are entitled to describing relevant background theory related to this thesis. For a thorough review and overview of the content, the reader is referred to Birge and Louveaux (2011); Shapiro et al. (2014), Bertsimas and Sim (2004), and Delage and Ye (2010) for theory about stochastic, robust, and distributionally robust optimization, respectively.

The remainder of this chapter is organized as follows. Section 2.1 presents the methodological background of decision-making problems. Section 2.2 describes the stochastic programming method, particularly, two-stage linear stochastic optimization. Section 2.3 focuses on robust optimization problems. Section 2.4 considers distributionally robust optimization problems. Finally, Section 2.5 presents the concluding remarks and summary of the chapter.

2.1. Methodological background

An optimization problem under uncertainty can be stated as the following standard optimization problem:

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \boldsymbol{\xi}), \tag{2.1}$$

where $\mathbf{x}$ is the vector of variables, $\boldsymbol{\xi}$ denotes the random vector of data parameters, and $\mathcal{X} \in \mathbb{R}^{\mathrm{d}}$ is a set (of nonnegative integers or binary integers variables) of feasible solutions not affected by uncertainty. Note that the cost function f is characterized as a minimization problem that depends on the random vector of parameters $\boldsymbol{\xi}$. Thus, $f(\cdot, \boldsymbol{\xi})$ is a random variable representing a family of optimization problems in each realization of $\boldsymbol{\xi}$. The goal is to find a unique optimal solution that minimizes the cost function f. This solution can be obtained according to the optimization approach, for instance, by computing the function expected cost, $\mathbb{E}\left(f(\mathbf{x}, \boldsymbol{\xi})\right)$, or by means of probability constraints, $\Pr\big(g(\mathbf{x}, \boldsymbol{\xi}) \leq 0\big) \geq 1 - \alpha$, to control the cost function risk to be satisfied with high probability, where $\alpha \in (0, 1)$ is the risk factor.

Based on the available information, different assumptions can be adopted to characterize the uncertain parameter $\boldsymbol{\xi}$ in problem (2.1). The characterization of the random vector, $\boldsymbol{\xi}$, in an optimization problem will depend on the uncertain data's assumptions. A deterministic approach, for instance, will assume expected values, $\bar{\boldsymbol{\xi}}$, of the data uncertainty based on expert judgment or historical data if available. Other optimization methodologies under uncertainty, represent unknown parameters assuming probabilistic estimates, partial or distributional information, such as TSO, robust optimization, and distributionally robust optimization, respectively. Regardless of the assumption made about the uncertain parameter, it is important to achieve the

most accurate estimate of the uncertain data based on the available information, to avoid unfavorable outcomes in the decision-making process.

In the rest of this chapter, we provide an overview of the methodologies mentioned above.

2.2. Stochastic Programming

Stochastic Programming (SP) is one of the most common approaches for modeling optimization problems under uncertainty. Different stochastic programming frameworks can be found in the literature. The most common approaches are two-stage linear optimization program (Birge & Louveaux, 2011), in which decisions are divided into stages, and Chance constraint linear programs involving probabilistic constraints (see, e.g., Charnes and Cooper (1959); Miller and Wagner (1965) and references therein). In this section and the rest of the chapter, we focus on stochastic linear programs of the form of two-stage stochastic optimization.

Below we specify the relevant notation and the probabilistic description of the uncertain parameters in two-stage stochastic programming based on Birge and Louveaux (2011).

The random experiments can be denoted by ω and the set of outcomes by Ω, so that $\omega \in \Omega$. The set $\mathcal{A}$ represents the collection of random events where $A \in \mathcal{A}$. A probability, $P(A)$, is associated to each event such that $0 \leq P(A) \leq 1$, $P(0) = 0$, $P(\Omega) = 1$. The probability space that describes the random events can be defined by $(\Omega, \mathcal{A}, P)$. Within the probability space the random parameter, $\boldsymbol{\xi}$, is defined as a particular random variable with a cumulative distribution $F_{\boldsymbol{\xi}}(x) = P(\omega|\boldsymbol{\xi} \leq x)$. The random variable can be defined as discrete or continuous. The discrete random variables take finite values within a countable range, $\boldsymbol{\xi}^{\mathrm{k}},\ k \in K$, described by a probability mass function in which,

$f(\xi^{\mathrm{k}}) = P(\boldsymbol{\xi} = \xi^{\mathrm{k}})$ such that $\sum_{k \in K} f(\xi^{\mathrm{k}}) = 1$, and its expectation is calculated by $\bar{\boldsymbol{\xi}} = \sum_{k \in K} \xi^{\mathrm{k}} f(\xi^{\mathrm{k}})$. The continuous random variables are represented by a probability density function, $f(\xi)$, and it can be defined by, $P(a \leq \boldsymbol{\xi} \leq b) = \int_a^{\mathrm{b}} f(\boldsymbol{\xi}) d\boldsymbol{\xi}$, in which the probability of $\boldsymbol{\xi}$ lays in the interval $[a, b]$. The expectation can be computed as $\bar{\boldsymbol{\xi}} = \int_{-\infty}^{\infty} \boldsymbol{\xi} dF(\xi)$, in which $F(\cdot)$ is the cumulative distribution.

2.2.1. Two-stage Stochastic Optimization

In decision-making problems under uncertainty, the decisions have to be made at different moments or stages in the time horizon, with incomplete information about the random parameters. At the upper stages, the available information is usually aggregated, but as the level of decision decrease, more detailed information is accessible. A common approach to represents this decision process in stages is Two-stage Stochastic Optimization (TSO), in which unknown data are represented as random variables (Birge & Louveaux, 2011).

TSO was first introduced by Dantzig (1955), and it assumes that the probability distribution of unknown parameters is known; thus, the decision-maker has full and accurate information to represent data uncertainty. The solutions of the TSO are provided through *recourse programs* in which some actions are taken after the uncertainty is revealed. The decision process is divided in two stages:

(i) *first-stage decisions.* These decisions are taken without knowledge of the uncertain realizations, and are commonly named, *here-and-now* decisions.

(ii) *second-stage decisions.* These decisions are corrective actions taken once the random data is revealed.

In order to characterize a TSO problem, we consider the optimization problem presented in (2.1) and assume that it is a stochastic linear program. We use boldfaced upper

and lower case symbols to represent vectors and matrices, respectively. The TSO can be represented in its compact form in the model (2.2), distinguishing between first-stage and second-stage decisions. The first term, $\boldsymbol{c}^{\intercal}\mathbf{x}$, corresponds to the first-stage, deterministic decisions. Here, the first-stage decisions are defined by the vector, $\mathbf{x}$, to which corresponds the vectors $\boldsymbol{c}$, and $\boldsymbol{b}$, and matrix $\boldsymbol{A}$, with the related dimensions. The second-stage term is the recourse function of the problem, $f(\mathbf{x}, \boldsymbol{\xi})$, which is represented as the expectation of the second-stage objective taken over all realizations of $\boldsymbol{\xi}$ within a given probability measure P.

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad & \boldsymbol{c}^{\intercal}\mathbf{x} + \mathbb{E}_P\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \\ \text{s.t.:} \quad & \boldsymbol{A}\mathbf{x} \leq \boldsymbol{b} \end{aligned} \tag{2.2}$$

The operational value function, $f(\mathbf{x}, \boldsymbol{\xi})$ can be represented as in the model in relation (2.3), where the second-stage decisions are defined by the vector $\mathbf{y}$. $\boldsymbol{h}$, $\boldsymbol{d}$, $\boldsymbol{W}$, and $\boldsymbol{B}$, corresponds to the vector and matrices, respectively, that become known when the random variable, $\boldsymbol{\xi}$, is realized. Each component of the vector and matrices can be a possible random variable.

$$\begin{aligned} f(\mathbf{x}, \boldsymbol{\xi}) = \min_{\mathbf{y}\geq 0} \quad & \boldsymbol{h}_{\boldsymbol{\xi}}^{\intercal}\mathbf{y} \\ \text{s.t.:} \quad & \boldsymbol{W}_{\boldsymbol{\xi}}\mathbf{x} + \boldsymbol{B}_{\boldsymbol{\xi}}\mathbf{y} \geq \boldsymbol{d}_{\boldsymbol{\xi}} \end{aligned} \tag{2.3}$$

The deterministic equivalent of the two-stage stochastic formulation is defined as the equations (2.4) (Birge & Louveaux, 2011). Let us define the second-stage decisions by the function $g_{\text{SP}}(\mathbf{x}) = \mathbb{E}_P[f(\mathbf{x}, \boldsymbol{\xi})]$. Then, first-stage decisions are taken by considering future events which are measured by the value function, $g_{\text{SP}}(\mathbf{x})$. We remark that this general formulation can be extended according to the specific problem. For instance, for problems considering first-stage or second-stage decisions as continuous and integer variables, the

domain can be replaced by $\mathbf{x} \geq 0$, and $\mathbf{y} \in \mathcal{Y}$, for first-stage and second-stage variables, respectively.

$$\begin{aligned} \textbf{SP}: \quad & \min_{\mathbf{x}\in\mathcal{X}} \quad \boldsymbol{c}^{\intercal}\mathbf{x} + g_{\text{SP}}(\mathbf{x}) \\ & \text{s.t.:} \quad \boldsymbol{A}\mathbf{x} \leq \boldsymbol{b} \\ & \text{where:} \quad g_{\text{SP}}(\mathbf{x}) = \mathbb{E}_P\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \end{aligned} \tag{2.4}$$

According to the type of problem, and which parameters are considered uncertain, there exist several formulation structures, in addition to the basic form of the two-stage stochastic program presented in equations (2.2). For instance, *complete recourse* problems, are of the form in equations (2.3), in which the matrix $\boldsymbol{B}$ is subject to uncertainty. Here, every first-stage solution, $\mathbf{x}$, under the constraints $\boldsymbol{A}\mathbf{x} \leq \boldsymbol{b}$ have a feasible completion in the second-stage (Birge & Louveaux, 2011). *Fixed recourse* problems, considers the recourse matrix, $\boldsymbol{B}$, as known. Such a formulation structure offers computational advantages because it allows us to characterize the feasibility region in a convenient way (Birge & Louveaux, 2011). There are also problems in which the parameter vector $\boldsymbol{h}$, and matrix $\boldsymbol{W}$ are fixed, and only the vector $\boldsymbol{d}$ is random. This formulation structure is named, *simple recourse*. This characterization allows the cost function, $g_{\text{SP}}(\mathbf{x})$, to be separable in the random components, $\boldsymbol{d}$, which is easy to solve.

Regardless of the formulation structure, stochastic optimization models should consider the *non-anticipativity conditions*. The decisions, $\mathbf{x}$, taken at the first stage, should be fixed for all future realizations in the second stage; this can be done by including non-anticipativity constraints in the formulation to enforce it. Further details about the properties of the aforementioned stochastic formulations can be found in Birge and Louveaux (2011).

2.2.2. Solution methods for Two-stage Stochastic Optimization

One of the main drawbacks of TSO problems is that its size (measured as the number of variables and constraints) can grow exponentially as the number of scenarios increases. Besides, the computation time can also be affected according to the problem modeling structure, in terms of the random variable's characterization, e.g., discrete or continuous. Thus, it is important to understand the problem structure in order to use a cost-efficient solution method.

Different approaches have been proposed in the literature, such as approximated techniques and decomposition-based methods. Approximated methods such as sampling, variance reduction, and scenario aggregation, are developed to approximate the optimal solution (Heitsch & Römisch, 2003; Homem-de Mello & Bayraksan, 2014). Decomposition methods, e.g., L-shaped, Benders, aims to find an exact or near-optimal solution of the linear program (Higle & Sen, 1991; Birge & Louveaux, 2011). Such methods are easily applicable if the model formulation includes complicating constraints, that can help to decompose the problem by scenarios (Conejo et al., 2010). Overall, linear problems can be usually solved by characterizing uncertainty with many scenarios, without affecting the solution time. In contrast, if the linear problem includes discrete variables, a decomposition technique is usually applied. For nonlinear problems involving continuous and discrete variables, decomposition techniques are also used (Conejo et al., 2010).

In this subsection, we cover the Sample Average Approximation approach, which lies within sampling solution methods in linear stochastic programs. For a broader detail of solution techniques based on decomposition methods, the reader is referred to Higle and Sen (1991), Conejo et al. (2006), and Birge and Louveaux (2011).

2.2.2.1. Sample Average Approximation solution method

Sample Average Approximation (SAA) method is a well-known approach in decision-making under uncertainty; it has been shown that under mild assumptions, it guarantees strong asymptotic performance and tractability for continuous and discrete distributions (see, e.g., Bertsimas et al. (2018b), Kleywegt et al. (2002)). The main drawback of SAA is that for a large number of samples and complex recourse functions, it can be computationally expensive to solve. Thus, the random parameters' sample size should represent the uncertainty, without affecting the resolution time. Hence, in some cases, scenario reduction techniques and sampling methods are advisable (Kleywegt et al., 2002).

SAA consists in replacing the recourse function, $g_{\text{SP}}(\mathbf{x})$, by a Monte Carlo estimate. The Monte Carlo method generates random samples of the uncertain parameter from the assumed continuous distribution density. The expected value function is, thus, approximated by the sample average function. The approximated problem is solved by deterministic optimization algorithms. Note that when assuming discrete probability distributions of the random parameters, the samples can be enumerated to compute the value function's expected value. On the contrary, when the probability distribution of the uncertain parameter is continuous, finding a solution for the TSO can be intractable; therefore, sampling techniques are commonly employed.

To characterize the SAA approach, let assume the random vector, $\boldsymbol{\xi}$, has finite support, in which ξ_k index the random samples,

$$g_{\text{SP}}^{\text{n}}(\mathbf{x}) = \sum_{k \in N} \frac{f(\mathbf{x}, \boldsymbol{\xi}_k)}{N}. \tag{2.5}$$

Then, for a two-stage stochastic model the SAA can be defined as follows:

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & c^{\intercal}\mathbf{x} + \frac{1}{N}\sum_{k\in N} \boldsymbol{h}_k^{\intercal}\mathbf{y}_k \\
\text{s.t.:} \quad & \boldsymbol{A}\mathbf{x} \leq \boldsymbol{b} \\
& \boldsymbol{W}_k\mathbf{x} + \boldsymbol{B}\mathbf{y}_k \geq d_k \quad \forall k \in N \\
& \mathbf{x} \in \mathcal{X}, \mathbf{y}_k \geq 0.
\end{aligned}
\tag{2.6}
$$

The samples, N, can be derived via Monte Carlo simulations or employing historical data. For a large size of N, the solutions to (2.6) will converge to an optimal solution (Birge & Louveaux, 2011). However, as stated before, scenario reduction techniques, e.g., probability distance method, are advisable to reduce the computation time while guaranteeing an accurate representation of uncertainty.

2.2.3. Quality metrics

TSO can become computationally difficult to solve according to the number of scenarios and the problem structure. In order to handle the computational difficulties, problems under uncertainty are usually solved by considering its deterministic version. Such programs consider expected values of the uncertain parameters, and therefore, are simpler to solve. Another alternative is to solve different deterministic problems, as independent scenarios and combine the solutions via heuristic methods (Birge & Louveaux, 2011).

Since stochastic programs are computationally expensive, it is important to assess the benefits and advantages of using stochastic programming, rather than a simpler and easier program, such as a deterministic approach. Two metrics are commonly employed to evaluate the accuracy and optimality of stochastic programs, namely, the Expected Value of Perfect Information (EVPI) and the Value of the Stochastic Solution (VSS).

Besides, *out-of-sample* techniques are also used to evaluate the outcome of the solutions by employing stochastic programming models.

In this subsection, we describe the metrics mentioned above, widely used to evaluate the solution performance of stochastic programs and problems under uncertainty.

A. Expected value of perfect information

The expected value of perfect information metric represents the value that a decision-maker would pay if he had perfect information about the future. Let us assume a TSO linear problem with fixed recourse, as in equations (2.7), in which the uncertainty is modeled through a finite set of scenarios. $\boldsymbol{\xi}$ is the random variable representing the scenarios within the set $\boldsymbol{\Xi}$ so that $\boldsymbol{\xi} \in \boldsymbol{\Xi}$. Here it is assumed that for all $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, there is at least one optimal solution.

$$\begin{aligned} &\min_{\mathbf{x}} z(\mathbf{x}, \boldsymbol{\xi}) = \boldsymbol{c}^{\intercal}\mathbf{x} + \min_{\mathbf{y}} \Big\{ \mathbf{h}_{\boldsymbol{\xi}}^{\intercal}\mathbf{y} | \boldsymbol{B}\mathbf{y} \geq \boldsymbol{d} - \boldsymbol{W}\mathbf{x}, \mathbf{y} \geq 0 \Big\} \\ &\text{s.t.:} \qquad\qquad \boldsymbol{A}\mathbf{x} \leq \boldsymbol{b},\ \mathbf{x} \in \mathcal{X} \end{aligned} \tag{2.7}$$

The optimal solution of problem (2.7) can be defined by $\bar{\mathbf{x}}(\boldsymbol{\xi})$. Assuming available perfect information, the objective function optimal value can be defined by $z(\bar{\mathbf{x}}(\boldsymbol{\xi}), \boldsymbol{\xi})$. Here, the objective value is obtained by relaxing the nonanticipativity constraints. This solution is commonly known as the *wait-and-see* solution, which we define as, Z^{K} in equations (2.8), where K stands for "known information".

$$Z^{\mathrm{K}} = \mathbb{E}_P \left[\min_{\mathbf{x}} z(\mathbf{x}, \boldsymbol{\xi}) \right] = \mathbb{E}_P\, z\big(\bar{\mathbf{x}}(\boldsymbol{\xi}), \boldsymbol{\xi}\big) \tag{2.8}$$

The solutions obtained from (2.8), can be compared with the *here-and-now* solutions defined as Z^{S}, and obtained by computing the expected value of z over all scenarios, $\boldsymbol{\xi}$, as in equation (2.9), in which $\mathbf{x}$, represents the optimal solution.

$$Z^{\mathrm{S}} = \min_{\mathbf{x}} \ \mathbb{E}_P \, z(\mathbf{x}, \boldsymbol{\xi}) \tag{2.9}$$

The EVPI is then computed as the difference between the here-and-now and wait-and-see solutions,

$$EVPI = Z^{\mathrm{S}} - Z^{\mathrm{K}}. \tag{2.10}$$

B. Value of the stochastic solution

The value of the stochastic solution metric assesses how much is gained from using stochastic programming instead of a deterministic approach. To compute the VSS, it is necessary to derive a new problem, obtained by replacing the random variables with their expected value, where the expectation of $\boldsymbol{\xi}$ is denoted by $\bar{\boldsymbol{\xi}} = \mathbb{E}(\boldsymbol{\xi})$. This new problem is called the *mean value problem*, Z^{M}, as represented in equation (2.11).

$$Z^{\mathrm{M}} = \min_{\mathbf{x}} z(\mathbf{x}, \bar{\boldsymbol{\xi}}) \tag{2.11}$$

The optimal solutions of Z^{M} in (2.11) can be defined by, $\bar{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})$, which are optimal values of the first-stage variables. Then, we can solve the stochastic programming problem by fixing the values of first-stage variables obtained by solving Z^{M}. This problem is decomposed by scenarios and easy to solve (Conejo et al., 2010). The expected result of using the Z^{M} solution, can be defined as,

$$Z^{\mathrm{M}^*} = \mathbb{E}_P \left(z(\bar{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}), \boldsymbol{\xi}) \right). \tag{2.12}$$

The Z^{M^*} problem in (2.12) measures the performance of $\bar{\mathrm{x}}(\bar{\boldsymbol{\xi}})$ when fixing its values into the first-stage variables, and obtaining second-stage solutions chosen optimally as a

function of $\bar{\mathbf{x}}(\bar{\boldsymbol{\xi}})$, and $\boldsymbol{\xi}$. For a minimization problem, the VSS can be computed as in (2.13), which is the difference between the here-and-now and expected value solutions,

$$VSS = Z^{\mathrm{M}^*} - Z^{\mathrm{S}}. \tag{2.13}$$

As a final remark, we would like to highlight the relationship between the values obtained from EVPI and VSS metrics, as indicated in Birge and Louveaux (2011). The two main properties that describe the metrics relation over uncertainty effects are listed below.

1. For any stochastic program, the metrics are nonnegative.
   $EVPI \geq 0$
   $VSS \geq 0$
2. For stochastic programs with fixed recourse matrix and objective coefficients, the metrics are bounded by the same quantity, and will be zero when the values of Z^{M^*} and Z^{M} are equal.
   $EVPI \leq Z^{\mathrm{M}^*} - Z^{\mathrm{M}}$
   $VSS \leq Z^{\mathrm{M}^*} - Z^{\mathrm{M}}$

C. Out-of-sample evaluation

In the stochastic programming field and overall optimization under uncertainty, other techniques can be used to assess the performance of stochastic models, such as *out-of-sample* evaluation. The procedure consists of simulating new scenarios (by fixing first-stage decisions), employing real data, or different distributional information to evaluate the reliability of the stochastic model's solutions. The robustness of the solutions can be evaluated via risk measures defined according to the model objective.

2.3. Robust Optimization

Robust Optimization (RO) is an alternative framework to stochastic optimization in which the uncertain parameter, $\boldsymbol{\xi}$, does not rely on probabilistic information; instead, uncertainty is modeled within an uncertainty set, $\mathcal{U}$. A generic robust optimization program takes the form:

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ \max_{\boldsymbol{\xi}\in\mathcal{U}} \big[f(\mathbf{x}, \boldsymbol{\xi}) \big] \right\}. \tag{2.14}$$

While in the TSO approach, the expected value of the objective function is optimized over a set of scenarios with known distributional information, in RO, a feasible and optimal solution is sought for the worst case of the objective function. RO approach is a useful methodology for applications in which infeasibility cannot be allowed since it provides an optimal solution that is feasible for any realization of the random data within the uncertainty set, i.e., the worst-case. The main drawback of RO is that it can lead to over-conservative solutions since it does not consider distributional information in case of available. However, the level of conservatism may be adjusted according to how the uncertainty set characterizes the uncertainty. Several robust approaches for linear optimization problems under uncertainty have been considered in the technical literature, differing according to the uncertainty set, $\mathcal{U}$, characterization.

The first contribution in RO was introduced by Soyster (1973), in which the uncertain parameters are modeled as variables that vary within an interval. Within this approach, the author obtained tractable robust optimization counterparts. In other words, under the Soyster approach, the uncertain optimization problem is reformulated into a deterministic convex program, guaranteeing feasibility for all realizations of the model uncertainties

(Goh & Sim, 2010). However, the interval modeling approach could lead to over-conservative solutions, since uncertain variables vary within a fixed range in the robust optimization counterpart; thus, the optimal solution tends to be at either end of the range. The solutions of this approach, also known as *box constrained uncertainty set*, give more value to robustness than to optimality, making it impractical for certain applications.

We remark that the Soyster (1973) approach considers *column-wise uncertainty*, defined as a specific set for each of the columns subject to uncertainty (Minoux, 2012). The box uncertainty set can be described as follows,

$$\mathcal{U}^{\text{box}} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^n : \boldsymbol{l} \leq \boldsymbol{\xi} \leq \boldsymbol{u} \right\}, \tag{2.15}$$

where the $\boldsymbol{\xi}$ is the random vector of decisions, that lies into a bounded and symmetric interval $[\boldsymbol{l}, \boldsymbol{u}]$, so that $\boldsymbol{l}$, $\boldsymbol{u} \in \mathbb{R}^n$ represent the lower and upper bound of the pre-defined interval, respectively, and $\boldsymbol{l} \leq \boldsymbol{u}$. Note that the uncertainty set, $\mathcal{U}^{\text{box}}$, assumes that all uncertain parameters vary simultaneously, i.e., every uncertain parameter takes the worst-value within the set, which may lead to over-conservative solutions.

In order to overcome the over-conservatism of the interval-based approach, the contributions Ben-Tal and Nemirovski (1998, 1999), El Ghaoui and Lebret (1997), and El Ghaoui et al. (1998) independently proposed a method based on ellipsoidal uncertainty sets for uncertain linear problems. The incorporation of ellipsoidal uncertainty allows the model to be less conservative, by considering correlation information in the uncertainty set. Here the uncertainty can be controlled by adjusting the size of the ellipsoidal sets. The main disadvantage of this approach is that it leads to nonlinear, although convex (Bertsimas & Sim, 2004), robust counterpart, which is more difficult to solve than the model proposed by Soyster (1973), especially in large scale applications.

The uncertainty set with ellipsoidal uncertainty can be described as follows,

$$\mathcal{U}^{\text{ell}} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^n : (\boldsymbol{\xi} - \bar{\boldsymbol{\xi}})^{\intercal} \, \boldsymbol{S}^{\intercal} \, (\boldsymbol{\xi} - \bar{\boldsymbol{\xi}}) \leq \Upsilon^2 \right\}, \tag{2.16}$$

where (as described before), $\boldsymbol{\xi}$ is the random vector of decisions that represents uncertainty. $\bar{\boldsymbol{\xi}}$ is the vector of the nominal (mean) value of the uncertain parameters, $\boldsymbol{S}$ represents a positive definite matrix of the relationship between the uncertain parameters (e.g., variance-covariance) and Υ is a safety parameter that determines the size of the ellipsoid to control the level of conservatism concerning the uncertainty. For values of $\Upsilon \geq 1$, more variability is acknowledged. The robust counterpart of (2.16) is an optimization over a quadratic constraint, which will result in a SOCP (Bertsimas et al., 2011).

Aiming to provide a trade-off between robustness and performance, Bertsimas and Sim (2004) proposed a framework based on a polyhedral uncertainty set. This approach allows for varying the degree of conservatism of the solution through a parameter called *budget of uncertainty*. As in Soyster (1973), the robust counterpart remains linear but can be generalized to discrete optimization problems. Besides, this new approach, in contrast to the box constrained uncertainty set method, allows controlling the level of conservatism for every constraint subject to uncertainty rather than simultaneously. Thus, the set has a row-wise structure, which allows that only a subset of the uncertain parameter changes to affect the solution (Bertsimas & Sim, 2004). The polyhedral uncertainty set can be described as follows,

$$\mathcal{U}^{\text{pol}} = \left\{ \boldsymbol{\xi} \in \mathbb{R}^n : |\xi_j - \bar{\xi_j}| \leq s_j, \, \sum_n \frac{|\xi_j - \bar{\xi_j}|}{s_j} \leq \Omega \right\}, \tag{2.17}$$

where n is the index set that contains the coefficients subject to uncertainty, s_j is the range of variation of coefficient j, and Ω is the adjustable parameter of robustness, also called *uncertainty budget*. The uncertainty budget is user-defined and indicates the number (not necessarily integer) of uncertain parameters that protect the solution against the level of conservatism so that $0 \leq \Omega \leq n$. Note that when $\Omega = 0$, no protection is considered, so the uncertain parameter is equivalent to its nominal value. Similarly, for $\Omega = |n|$ it becomes the Soyster (1973) method in which full protection is acknowledged.

From the mathematical modeling point of view, the robust optimization approaches differ according to problem structure, namely, *single-stage* and *adjustable* robust optimization. In the former approach, the decisions are not adjusted once the uncertainty is realized; thus, recourse actions cannot be taken in response to changes in the wait-and-see decisions. Some examples of single-stage works are the contributions previously presented, Soyster (1973); Ben-Tal and Nemirovski (1998); El Ghaoui and Lebret (1997); Bertsimas and Sim (2004). Adjustable robust linear programs, similar to TSO problems, consider adjustable decisions or recourse actions in the second-stage problem. The *two-stage* robust decision framework was addressed by Ben-Tal et al. (2004) in which the decisions are divided into stages, guaranteeing to some extent consistency in the decision-making. This approach, in particular, is explained in the next section, which is the focus of this thesis. For further review of the aforementioned robust optimization methods, the reader is referred to Ben-Tal and Nemirovski (1999), Bertsimas et al. (2011), Gorissen et al. (2015) and Sözüer and Thiele (2016).

2.3.1. Two-stage Robust Optimization

Adjustable robust optimization (RO henceforth) was introduced in Ben-Tal et al. (2004). The authors studied the problem in which decisions are flexible to adapt to changes in the uncertain parameters. Thus, recourse actions can be taken in response to future

variations in decisions. In Ben-Tal et al. (2004), it was showed that the adjustable RO is in general NP-hard. The adjustable RO framework is similar to stochastic programming with recourse, in which several stages can be considered according to how the decision process is conformed.

In particular, two-stage robust optimization considers two stages; first-stage decisions are taken before the realization of uncertainty, and second-stage decisions involve recourse or correction actions made once the uncertain events are realized. Unlike stochastic programming in which uncertainty is assumed to take values within a set of scenarios with a known probability distribution, in RO, decisions belong to an uncertainty set, that can be defined as we detailed in Subsection 2.3.

To illustrate the adjustable RO approach in two stages, let us recall the compact linear optimization problem presented in equations (2.2)–(2.3), in which we defined the operational value function $f(\mathbf{x}, \boldsymbol{\xi})$,

$$f(\mathbf{x}, \boldsymbol{\xi}) = \min_{\mathbf{y} \geq 0} \quad \Big\{ \boldsymbol{h}_{\boldsymbol{\xi}}^{\intercal} \mathbf{y} \mid \boldsymbol{W}_{\boldsymbol{\xi}} \mathbf{x} + \boldsymbol{B}_{\boldsymbol{\xi}} \mathbf{y} \geq d_{\boldsymbol{\xi}} \Big\}. \tag{2.18}$$

In equation (2.19), instead of obtaining the expectation over $f(\mathbf{x}, \boldsymbol{\xi})$ as in (2.2), we compute the worst-case operational cost under first-stage decisions, $\boldsymbol{x}$, and the observed vector of uncertainty, $\boldsymbol{\xi} \in \mathcal{U}$, defined by the cost function $g_{\text{RO}}(\mathbf{x}) = \max_{\boldsymbol{\xi} \in \mathcal{U}} [f(\mathbf{x}, \boldsymbol{\xi})]$.

The two-stage robust counterpart can be described as follows,

$$\begin{aligned} \textbf{RO}: \qquad & \min_{\mathbf{x} \in \mathcal{X}} \quad \boldsymbol{c}^{\intercal} \mathbf{x} + g_{\text{RO}}(\mathbf{x}) \\ \text{where:} \qquad & g_{\text{RO}}(\mathbf{x}) = \max_{\boldsymbol{\xi} \in \mathcal{U}} \Big[f(\mathbf{x}, \boldsymbol{\xi}) \Big]. \end{aligned} \tag{2.19}$$

The robust counterpart model in (2.19) aims to find a feasible solution under any realization of the uncertain parameter, $\boldsymbol{\xi}$, within the pre-specified uncertainty set, $\mathcal{U}$.

2.3.2. Solution methods for Two-stage Robust Optimization

Adaptive robust optimization models are difficult to compute; they are a tri-level optimization model that has been shown to be NP-hard (Ben-Tal et al., 2004). Several solutions strategies have been studied in the literature, aiming to find tractable methodologies to overcome the computational difficulty, for instance, approximation algorithms, decomposition-based methods, and scenario-based approaches. We remark that we limit this subsection to solution strategies applicable to two-stage robust counterpart models with continuous recourse variables, which is the scope of this thesis. For further detail of solution methodologies involving discrete recourse variables, the reader is referred to Zhao and Zeng (2012a).

Approximation algorithms were studied in Ben-Tal et al. (2004). The authors proposed the use of affine decision rules, named as *affinely adjustable robust counterpart* approach, in which the recourse decisions are restricted to being affinely dependent on the uncertain parameters. The main assumption is that the problem has a fixed recourse scheme (i.e., the recourse matrix, $\boldsymbol{B}$, in (2.18) is known). The optimality of the linear decision rule method was later studied in Bertsimas, Iancu, and Parrilo (2010) to prove the solution approach's tractability. For the case of discrete second-stage variables, Bertsimas and Caramanis (2010) introduced the *finite adaptability* approach, in which discrete variables are modeled as a piecewise constant function of the uncertainty. For large-scale problems, such approximation techniques are generally NP-hard, since the computational time grows with the number of partitions.

Decomposition-based methods, such as Benders decomposition (Benders, 1962) and column-and-constraint generation (CCG) algorithms, are used for large scale linear optimization problems. Rather than providing near-optimal solutions as in the fixed-recourse framework, global optimality may be obtained in a short time (depending on

the problem structure). The general idea behind the decomposition methods algorithms is to dynamically generate new solutions in a master-subproblem framework through an iterative process that provides bounds of the optimal solution (González Cobos, 2019). Applications of the Bender decomposition algorithm for two-stage robust models can be found in Zhao and Zeng (2012b). Here, the value function of the first-stage decisions is constructed by employing dual solutions of the second-stage decisions. The CCG algorithm, similar to Benders, creates cutting planes but with primal decision variables (Zeng & Zhao, 2013). Overall, it has been proven that for recourse problems involving continuous variables, both algorithms converge to an optimal solution in finite iterations (Bertsimas et al., 2012; Zhao & Zeng, 2012a; Gabrel et al., 2014).

The uncertainty in robust models can also be characterized by scenarios. Recent studies in the literature have considered this approach defined as *data driven robust optimization*, in which data is employed to design the uncertainty set, see, e.g., Bertsimas et al. (2018a), Bertsimas et al. (2018b), and Bertsimas and Kallus (2020). Such scenario-based methods are particularly suitable for robust approaches under polyhedral uncertainty sets and continuous recourse variables. The procedure is similar to the SAA approach described in Subsection 2.2.2.1; the second-stage is replaced with the uncertain constraints related to the scenarios representing the vertexes of the polyhedron that characterize the uncertainty (Bertsimas et al., 2018a).

A tractable reformulation duality-based of problem (2.19) will be derived in Subsection (2.4.2.1), by considering a box uncertainty set.

2.4. Distributionally Robust Optimization

The Distributionally Robust Optimization (DRO) approach is known to be a generalization of the stochastic and robust optimization approaches in which limited

distributional information of the uncertain parameter, $\boldsymbol{\xi}$, is considered. In contrast to stochastic programming in which the uncertainty is based on knowledge of the probability distribution, the distributionally robust optimization approach stands on the paradigm of *ambiguity* in decision-making (Bertsimas, Sim, & Zhang, 2019). A generic distributionally robust optimization program takes the form:

$$\min_{\mathbf{x}\in\mathcal{X}} \left\{ \sup_{P\in\mathcal{D}} \mathbb{E}_P \left[f(\mathbf{x}, \boldsymbol{\xi}) \right] \right\}, \tag{2.20}$$

where $\mathcal{D}$ is the ambiguity set that represents the family of probability distributions that characterize the random parameters. Hence, a DRO model seeks to hedge against the worst-case probability distribution within a pre-defined ambiguity set, $\mathcal{D}$. The definition of $\mathcal{D}$ is crucial in the robust optimization setting since it defines the computational tractability of the model and the degree of conservatism of the solution.

The adoption of limited distributional information in stochastic programming was first considered in Scarf (1958). Scarf studied a minimax stochastic problem applied to inventory planning under known moment information (first and second-order) of demand. It was shown that the proposed model could be reformulated as a tractable optimization problem. A similar approach was also studied in Žáčková (1966), assuming knowledge of the mean and support of the uncertain variables. Additional works to this minimax stochastic programming approach can be found in Dupačová (1987), Breton and El Hachem (1995), and Shapiro and Ahmed (2004).

More recently, the DRO approach has gained considerable interest in a number of applications, including revenue management, portfolio management, scheduling, and power system management. The studies can be subdivided into three main streams according to the ambiguity characterization, namely, (i) confidence region of the goodness of fit, (ii) probability distance (e.g., Kullback–Leibler, Wasserstein metric), and (iii)

moment information. We remark that this classification includes the most common approaches found in the literature; various ambiguity sets have been studied beyond this classification. Refer to Ben-Tal et al. (2013), Hanasusanto et al. (2015), and Postek et al. (2016) for a broad description of the DRO approach. Below we briefly detail the aforementioned approaches.

Goodness-of-Fit. The uncertainty set under the Goodness-of-Fit (GoF) approach contains all distributions that pass prescribed statistical tests (Bertsimas et al., 2018b), e.g., Kolmogorov-Smirnov or Anderson-Darling for univariate distributions. The GoF approach was employed in Bertsimas et al. (2018b) to derive a *robust SAA* (termed by authors), applied to a data-driven DRO by means of statistical hypothesis testing. The authors proposed an enhanced SAA guaranteeing asymptotic convergence and tractability for a finite sample, which may be unstable in out-of-sample-evaluation. The proposed robust SAA procedure consists of employing a GoF test at a significance level, $0 < \alpha < 1$, to construct the uncertainty set that contains a true probability distribution within the confidence region at least $1 - \alpha$. Similarly, for a given uncertainty set containing the true distribution at a significance level of at least $1 - \alpha$ (as a reference of the sampling distribution), a GoF can be constructed with a significance level of α.

A general representation of the ambiguity set under GoF has the form: $\mathcal{D}^{\text{GoF}} = \{P^{\alpha}_{S_N} \ : \ S_N(P_0, \xi^1, ..., \xi^N)\}$, where $\mathcal{D}^{\text{GoF}}$ represents the ambiguity set considering a GoF approach, S_N indicates a particular statistic test, which takes values $\xi^1, ..., \xi^N$ denoting independent and identically distributed (iid) data of dimension N, and P_0 is the *hypothetical* distribution. The confidence region (i.e., set of all distributions, P_0, that past the test) is denoted by, $P^{\alpha}_{S_N}(\xi^1, ..., \xi^N) = \{P_0 \in \mathcal{P}(\Xi) : S_N(P_0, \xi^1, ..., \xi^N) \leq Q_{S_N}(\alpha)\}$, such that, Ξ is the support (closed) of $\boldsymbol{\xi}$, denoted by the set of probability distributions over $\mathcal{P}(\Xi)$, Q_{S_N} represents a threshold which depend on the true distribution for a defined

confidence level, α. The null hypothesis is rejected if $P_0 > Q_{S_N}$. Overall, the GoF approach accounts for the evaluation of the entire distribution via its confidence region instead of considering only moments (mean and covariance) or the probability distribution distance.

Probability distance. This approach defines the ambiguity set as a ball in the space of probability distributions; thus, the degree of conservatism can be controlled by adjusting its radius. The ambiguity set under probability distance contains a family of distributions close to the nominal distribution related to the evaluated metric (Esfahani & Kuhn, 2018).

A general representation of the ambiguity set under probability distance has the form: $\mathcal{D}^{\text{dist}} = \{P : d(P, \widehat{P}) \leq \beta\}$, where $\mathcal{D}^{\text{dist}}$ is a bounded ambiguity set, $P \in \mathcal{D}$ denotes probability measures defined under the same sample space, $d(\cdot, \cdot)$ is a distance function between probability measures, $\widehat{P}$ denotes a central (or reference) probability measure, and $\beta \geq 0$ is a given constant. Some of the most common measures studied in the literature are the Prohorov metric, Kullback–Leibler, and the Wasserstein metric.

The *Prohorov metric* (Billingsley, 2013) is important in probability and statistics theory because it metrizises the weak convergence between probability measures. The authors in Erdoğan and Iyengar (2006) employed this metric for the study of ambiguous chance-constrained problems under a sample-based approximation method. The uncertainty set is defined by norm balls described in terms of the Prohorov metric.

The *Kullback–Leibler* (KL) divergence measure also named *relative entropy*, belongs to the class of distances called *phi-divergence*; see, e.g., Ben-Tal et al. (2013) and references therein. It compares the difference between two probability distributions, assuming known, finite and discrete support. This metric is known to be asymmetric and, therefore, is not a metric since it doesn't satisfy the triangle inequality. The KL measure has been mostly applied to chance-constrained problems providing tractable

approximations for the ambiguous chance constraint, for instance, Calafiore and El Ghaoui (2006); Hu et al. (2013); Hu and Hong (2013) and, Jiang and Guan (2016).

The *Wasserstein metric* (Kantorovich & Rubinstein, 1958), has been widely used for modeling distributional ambiguity; see, e.g., Pflug and Wozabal (2007), Wozabal (2012), Mehrotra and Zhang (2014), Gao and Kleywegt (2016), and Esfahani and Kuhn (2018). The Wasserstein distance (also referred to as *Kantorovich metric*), of two distributions, can be defined as the transportation cost for moving the probability mass between both distributions. In other words, the Wasserstein distance describes the cost of an optimal mass transportation plan (Esfahani & Kuhn, 2018). The distributionally robust optimization problems under Wasserstein/Kantorovich metric are known to be difficult to solve. Most existing methods to solve such problems rely on global optimization techniques (Esfahani & Kuhn, 2018). In this regard, the authors in Esfahani and Kuhn (2018) studied the tractability issue of the worst-case expectation over a Wasserstein ball for a data-driven distributionally robust model. The authors demonstrated the tractability of the Wasserstein distance that can be achieved by transforming the inner problem into a finite-dimensional convex program. Besides, it was shown that for a finite sample, such an approach provides good out-of-samples performance. The authors also pointed out that moment-based ambiguity set approaches provide advantages over the Wasserstein distance metric in terms of tractability properties. Below we describe the moment-based ambiguity set approach, which is within the scope of this thesis.

Moment information. This approach assumes that only moment information (such as mean, covariance, support) of the distribution of the uncertain parameters is known (estimated) to the decision-maker. Several definitions of ambiguity set under moment information have been proposed in the literature.

The first study based on moment information is found in Scarf (1958). The ambiguity set is defined by considering *linear constraints* on the first and second moments of the demand distribution for a (one-stage) single-product news-vendor problem. The solution method relies on characterizing the worst-case distribution as a point distribution. Thereafter a similar linear constraint-based approach has been applied to several studies; see, e.g., Dupačová (1987); Bertsimas and Popescu (2005), considering the knowledge of mean and support of the uncertain parameters, and Yue et al. (2006); Zhu and Fukushima (2009); Prékopa (2013), imposing exact knowledge on mean and covariance matrix constraints.

A different approach to describe the uncertainty set under moment information is to consider *conic constraints*; see, e.g., Bertsimas, Doan, et al. (2010); Delage and Ye (2010); Wiesemann et al. (2014). More specifically, Delage and Ye (2010) propose a general uncertainty set, assuming knowledge of the distribution's support, mean and second-moment matrix, which lied in a confidence region rather than considering point estimation as in previous studies. The authors show that due to the structure of this new generalized uncertainty set, a DRO model can be solved in polynomial time (through an ellipsoid method), offering computational advantages over previously reported approaches in the literature. The proposed DRO model is developed under a data-driven approach and applied to a portfolio selection problem.

Mehrotra and Zhang (2014), extended the study in Delage and Ye (2010) for least-square problems. The authors considered three uncertainty sets, defined by moment constraints, norm bounds, and confidence regions. The model is solved in polynomial time by using a semidefinite programming method. In Mehrotra and Papp (2014) high-dimensional moments (e.g., third and fourth marginal) are considered to define the uncertainty set, which is able to provide the overall shape of the distribution. However,

the proposed cutting surface method based on a semi-infinite programming approach does not guarantee a polynomial solution. More recently, a generalized ambiguity set based on conic constraints is proposed in Wiesemann et al. (2014). The unified framework considers several uncertainties sets previously presented in the literature. Tractability conditions for the proposed approach are presented.

Motivated by the new emerging of data availability in various fields, there exists a vast literature in data-driven distributionally robust optimization; see, e.g., Calafiore and El Ghaoui (2006), Delage and Ye (2010), Jiang and Guan (2016), Esfahani and Kuhn (2018), Bertsimas et al. (2018b), and Bertsimas et al. (2018a). In this setting, the main objective is to construct the uncertainty set based on historical data. For instance, Delage and Ye (2010), design the ambiguity set from historical data by employing the McDiarmid's inequality (Mcdiarmid et al., 1998), to define the confidence region of the mean and covariance matrix of the uncertain parameter. In contrast, the authors in Bertsimas et al. (2018a) develop general uncertainty sets by employing a hypothesis test. It was shown that this framework leads to less conservative solutions while maintaining similar probabilistic guarantees and enhanced robustness properties compared to the previous approaches in the literature. For a broad review on uncertainty set characterization in distributionally robust optimization the reader is referred to Gabrel et al. (2014), Wiesemann et al. (2014), and Sözüer and Thiele (2016).

Below we summarize the most common formulations of the ambiguity set, $\mathcal{D}^{\text{mom}}$, under moment information to solve a problem of type (2.20).

1. *Considering all distributions with exact mean vector, $\bar{\boldsymbol{\xi}}$, and covariance matrix, $\boldsymbol{\Sigma}$, constraints*, (Scarf, 1958; Ghaoui et al., 2003; Yue et al., 2006; Popescu, 2007):

$$\mathcal{D}_1^{\text{mom}} = \left\{ P \in \mathcal{P}^{\text{n}} \;\middle|\; \begin{array}{l} P\{\boldsymbol{\xi} \in \Xi\} = 1 \\ \mathbb{E}_P[\boldsymbol{\xi}] = \bar{\boldsymbol{\xi}} \\ \mathbb{E}_P\left[(\boldsymbol{\xi} - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi} - \bar{\boldsymbol{\xi}})^{\intercal}\right] = \boldsymbol{\Sigma} \end{array} \right\}, \tag{2.21}$$

where $\mathcal{P}^{\text{n}}$ represents the set of all probability measures on $\mathbb{R}^n$, $\bar{\boldsymbol{\xi}}$ stands for the mean vector, and $\boldsymbol{\Sigma}$ is the covariance matrix.

2. *Considering all distributions with exact mean vector, $\bar{\boldsymbol{\xi}}$, and support, $\boldsymbol{\Xi}$, constraints*, (Dupačová, 1987; Bertsimas & Popescu, 2005):

$$\mathcal{D}_2^{\text{mom}} = \left\{ P \in \mathcal{P}^{\text{n}} \;\middle|\; \begin{array}{l} P\{\boldsymbol{\xi} \in \Xi\} = 1 \\ \mathbb{E}_P[\boldsymbol{\xi}] = \bar{\boldsymbol{\xi}} \end{array} \right\}, \tag{2.22}$$

where Ξ represents the support set of the random vector $\boldsymbol{\xi}$, and $P\{\boldsymbol{\xi} \in \Xi\} = 1$ guarantees that the uncertainty realizations are constrained within the support set.

3. *Considering all distributions with known support, Ξ, ambiguous mean vector, $\bar{\boldsymbol{\xi}}$ and ambiguous covariance matrix, $\boldsymbol{\Sigma}$*, constraints, (Delage & Ye, 2010):

$$\mathcal{D}_3^{\text{mom}} = \left\{ P \in \mathcal{P}^{\text{n}} \;\middle|\; \begin{array}{l} P\{\boldsymbol{\xi} \in \Xi\} = 1 \\ \left(\mathbb{E}_P[\boldsymbol{\xi}] - \bar{\boldsymbol{\xi}}\right)^{\intercal} \boldsymbol{\Sigma}^{-1} \left(\mathbb{E}_P[\boldsymbol{\xi}] - \bar{\boldsymbol{\xi}}\right) \leq \gamma_1 \\ \mathbb{E}_P\left[(\boldsymbol{\xi} - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi} - \bar{\boldsymbol{\xi}})^{\intercal}\right] \leq \gamma_2 \, \boldsymbol{\Sigma} \end{array} \right\}. \tag{2.23}$$

The generalized ambiguity set, (2.23), proposed by Delage and Ye (2010), aims to control the estimation errors on the mean vector and covariance matrix. It considers the parameters, $\gamma_1 \geq 0$ and $\gamma_2 \geq 1$, which define the size of the ambiguity set.

The DRO has been studied within a multi-stage decision framework (see, for instance, Bertsimas and Goyal (2010); Bansal et al. (2018); Bertsimas, Sim, and Zhang (2019) and references therein). Similar to how we explained in previous sections, the framework divides the decision-making process into stages and is termed in the technical literature as the *"Adaptive distributionally robust optimization problem"*. In Subsection 2.4.1 we focus on the two-stage distributionally robust linear optimization problem, following a similar criteria of Subsections 2.2.1 and 2.3.1. We consider moment information to characterize the uncertainty set, and to derive a tractable solution methodology.

2.4.1. Two-stage Distributionally Robust Optimization

The two-stage DRO optimization approach, follows the same decision criteria of two-stage stochastic (Birge & Louveaux, 2011) and robust optimization (Bertsimas et al., 2012) frameworks. The decisions are divided into two stages, where the first-stage or *here-and-know* decisions are taken before the realization of uncertainty, and the second-stage *wait-and-see* or recourse decisions are made once the uncertain events are realized. Thus, decisions can be adjusted in response to changes in future realizations. Within two-stage DRO framework several classes of optimization problems are studied in the literature which differ according to the model structure (e.g, LP, MILP, MIP) and the definition of the ambiguity set. Some interesting contributions within this framework of decisions can be found in, Goh and Sim (2010), Bansal et al. (2018), Shang and You (2018), Bertsimas, Sim, and Zhang (2019), Y. Wang et al. (2019), and Velloso et al. (2020). In the following subsections, we focus on two-stage distributionally robust linear programs under moment information.

Based on the generic distributionally robust program in equation (2.20) we further consider the following two-stage DRO problem:

$$\begin{aligned} \textbf{DRO:} \quad & \min_{\mathbf{x}\in\mathcal{X}} \quad \boldsymbol{c}^{\intercal}\mathbf{x} + g_{\text{DRO}}(\mathbf{x}) \\ \text{where:} \quad & g_{\text{DRO}}(\mathbf{x}) = \sup_{P\in\mathcal{D}^{\text{mom}}} \mathbb{E}_P\Big[f(\mathbf{x},\boldsymbol{\xi})\Big]. \end{aligned} \tag{2.24}$$

The goal of problem (2.24) is to determine the *first-stage* decisions, $\mathbf{x}$, (which can be binary or continuous variables), by minimizing the sum of the first-stage cost and an ambiguity-averse expectation measure, $g_{\text{DRO}}(\mathbf{x})$, (Bertsimas, Sim, & Zhang, 2019). The *second-stage* recourse function $g_{\text{DRO}}(\mathbf{x})$ computes the worst-case expected cost of all probability distributions over the ambiguity set $\mathcal{D}^{\text{mom}}$.

Without loss of generality, we assume that function $f(\mathbf{x},\boldsymbol{\xi})$ is a linear optimization problem, defined on the first-stage (deterministic) decisions, $\mathbf{x}$, and the uncertain realizations, $\boldsymbol{\xi}$. Here the *second-stage* decision is represented by the vector $\mathbf{y}$. The inner linear optimization problem can be defined as follows:

$$\begin{aligned} f(\mathbf{x},\boldsymbol{\xi}) = \min_{\mathbf{y}\geq 0} \quad & \boldsymbol{h}^{\intercal}\mathbf{y} \\ \text{s.t.:} \quad & \boldsymbol{B}\mathbf{y} \geq d_{\boldsymbol{\xi}} - \boldsymbol{W}\mathbf{x}, \end{aligned} \tag{2.25}$$

where the uncertain parameter, d, of the second-stage problem is on the right-hand side of the uncertain constraint. Therefore, $f(\mathbf{x},\boldsymbol{\xi})$ can be defined as a convex function on $\mathbf{x}$ and $\boldsymbol{\xi}$. We remark that in the context of stochastic programming, problem (2.24) is categorized as *simple recourse*, in which the parameter vector $\boldsymbol{h}$, and matrix $\boldsymbol{W}$ are fixed, and only the vector d is random (Birge & Louveaux, 2011).

As we explained in the previous section depending on the available information about $\boldsymbol{\xi}$, the uncertainty can be characterized according to the definition of the ambiguity set. We consider the ambiguity set, $\mathcal{D}_2^{\text{mom}}$, defined in equation (2.22), which assume knowledge of

the mean vector, $\bar{\boldsymbol{\xi}}$, and support, Ξ, of the uncertain parameter. Then, the ambiguity-averse operational cost function, $g_{\text{DRO}}(\mathbf{x})$, can be expressed as an infinite dimensional linear optimization problem in equation (2.26):

$$\begin{aligned} g_{\text{DRO}}(\mathbf{x}) = \sup_{P \in \mathcal{D}^{\text{mom}}} & \int_{\Xi} f(\mathbf{x}, \boldsymbol{\xi})\, dP(\boldsymbol{\xi}), \\ \text{s.t.:} & \int_{\Xi} dP(\boldsymbol{\xi}) = 1 \qquad : \alpha \\ & \int_{\Xi} \boldsymbol{\xi} dP(\boldsymbol{\xi}) = \bar{\boldsymbol{\xi}} \qquad : \boldsymbol{\gamma} \end{aligned} \tag{2.26}$$

where the symbols, α and $\boldsymbol{\gamma}$, are the dual variables related to moment problem constraints. Note that under this formulation, (2.24) is a min-sup problem that cannot be solved with commercial solvers.

In the next subsection, we describe the up-to-date solutions methodologies to solve two-stage distributionally robust linear optimization problems for ambiguity set constructions under moment information. Besides, we derive a tractable methodology similar to Pozo et al. (2018) to solve problem (2.24) based on duality theory.

2.4.2. Solution methods for Two-stage Distributionally Robust Optimization

Adaptive DRO models are usually hard to solve. Accordingly, Bertsimas, Doan, et al. (2010) showed that such problems could be NP-hard for right-hand side uncertainty models. Several solution methods have been proposed in the literature to solve adaptive distributional robust optimization models, including exact and approximated approaches. Approximated methods (e.g., linear decision rule) are commonly employed to solve polynomial sized problems that can be computationally expensive to be solved by exact methods.

Besides, the solution methods differ according to the ambiguity set characterization and the structure of the recourse function $f(\mathbf{x}, \boldsymbol{\xi})$. For instance, when choosing an uncertainty set of the type in equations (2.23), the recourse function involves quadratic constraints, so that, its dual will lead to SOCP. For the case of considering first moments and support (e.g., polyhedral uncertainty) as in (2.22), the reformulation will result in an equivalent linear optimization problem. Of course, this is valid for the case in which the subproblem involves continuous variables so that its dual can be obtained (Bertsimas et al., 2011).

We briefly detail solution methods in the literature for adaptive DRO models under moment information. The reader is referred to Goh and Sim (2010) and Kuhn et al. (2011) for an overview of tractable reformulations to generic two-stage and multi-stage distributional linear robust programs.

Exact methods to solve adaptive DRO problems typically rely on duality theory approaches to find an equivalent finite formulation for the recourse function. This problem can then be solved by a vertex enumeration approach (i.e., including all the extreme points of the feasible set), and considering scenarios that represent the uncertainty described by the ambiguity set, see, e.g., Pozo et al. (2018). The optimization problem's size can increase exponentially according to the uncertain parameter dimension, which can be computationally intractable for large size problems. Thus, decomposition algorithms are employed, aiming to find tractable reformulations. For instance, a decomposition method based on Bender's algorithm and an L-shaped method is employed in Bansal et al. (2018) to solve two-stage DRO problems with first-stage binary variables and mixed binary variables in the second-stage. The problem, which is based on moments and Kantorovich ambiguity sets resulted in optimal solutions obtained in a short time.

Recently, an enhanced CCG algorithm is developed in Velloso et al. (2020) for solving two-stage DRO models with right-hand side uncertainty and continuous second-stage variables under first-moment information. In the traditional CCG algorithm usually applied in RO, the iterative process between the master and subproblem is based on a single scenario (at each iteration) to compute the worst-case value of the recourse function. In contrast, the enhanced CCG proposed by Velloso et al. (2020) considers that since in adaptive DRO, the recourse function relies on the worst-case *expected* value, more scenarios are needed to compute the mean value. Thus, an inner loop is included based on a Dantzig-Wolfe procedure to obtain the expected value of the recourse function, considering more scenarios at each iteration of the CCG algorithm. Better solutions were reported compared to the traditional CCG method.

Approximated algorithms are used to derive tractable reformulations for high dimensional problems. A common approach is the Linear Decision rule (LDR) technique also studied in robust optimization (Ben-Tal et al., 2004) as well as stochastic optimization (Kuhn et al., 2011). This approximated method consists in assuming that recourse decisions are affinely dependent on the uncertain parameters. According to Ben-Tal et al. (2004), performance improvement can be obtained under this approach for certain types of problems.

Lately researchers have been considering the LDR technique in DRO, see, e.g., Chen et al. (2007), Goh and Sim (2010), S. Wang et al. (2017) and Bertsimas, Sim, and Zhang (2019). In particular, Bertsimas, Sim, and Zhang (2019) developed a tractable formulation for adaptive DRO problems by considering a new variant of the LDR technique under a lifted ambiguity set. A scalable framework is proposed for SOC ambiguity sets; defined as *partial cross-moment* ambiguity set. In contrast to the previous studies that consider marginal moment ambiguity sets, the authors showed that by employing cross-moments,

the solutions are less conservative and yield tractable models. To guarantee feasibility of the second-stage problem, Bertsimas, Sim, and Zhang assumes that the recourse matrix (i.e., $\boldsymbol{B}$) and the cost vector (i.e., $\boldsymbol{h}$) of the recourse function, $f(\mathbf{x}, \boldsymbol{\xi})$, are constants and that it has *relatively complete recourse*, i.e., for every first-stage solution and uncertain realization, the second-stage is feasible (Birge & Louveaux, 2011).

Another approach is based on semidefinite programming (SDP) by employing a moment decomposition approach. This method is suitable for problems in which first and second moments characterize the ambiguity set. Kong et al. (2013) proposes an approximation algorithm based on the conic representation of the recourse function. A cross-moment ambiguity set is considered characterized by partial distributional information of the mean and covariance matrix of the uncertain parameter. A copositive cone programming reformulation is obtained, and a semidefinite program is used to approximate the solutions.

In the next subsection, we present an exact solution method to solve problem (2.24).

2.4.2.1. Scenario-based equivalent finite formulation

The model in the set of equations (2.26) is a linear optimization problem in the space of distributions with an infinite number of variables, each one associated with a probability element, $dP(\boldsymbol{\xi})$. In order to find a finite single-level equivalent formulation we consider a scenario-based approach to reformulate the model into an equivalent finite mixed-integer linear program. A dual formulation is developed that results in a problem with infinitely many linear constraints, one for each scenario, $\boldsymbol{\xi} \in \Xi$, and it can be defined as follows:

$$\begin{aligned} g_{\text{DRO}}(\mathbf{x}) = \min_{\alpha,\gamma} \quad & \alpha + \boldsymbol{\gamma}^{\intercal}\bar{\boldsymbol{\xi}} \\ \text{s.t.:} \quad & f(\boldsymbol{x}, \boldsymbol{\xi}) \leq \alpha + \boldsymbol{\gamma}^{\intercal}\boldsymbol{\xi} \qquad \forall \boldsymbol{\xi} \in \Xi. \end{aligned} \tag{2.27}$$

In order to reformulate problem (2.27) into a fine-dimensional one, we remark that the infinite set of linear constraints in problem (2.27) can be represented as,

$$\max_{\boldsymbol{\xi}\in\Xi} \left\{ f(\mathbf{x},\boldsymbol{\xi}) - \boldsymbol{\gamma}^{\intercal}\boldsymbol{\xi} \right\} \leq \alpha. \tag{2.28}$$

By definition, $f(\mathbf{x},\boldsymbol{\xi})$ is a convex function on $\boldsymbol{\xi}$. Thus, expression $f(\mathbf{x},\boldsymbol{\xi}) - \boldsymbol{\gamma}^{\intercal}\boldsymbol{\xi}$ is a convex function on $\boldsymbol{\xi}$. Therefore, the optimal value of $\max_{\boldsymbol{\xi}\in\Xi} \left\{ f(\mathbf{x},\boldsymbol{\xi}) - \boldsymbol{\gamma}^{\intercal}\boldsymbol{\xi} \right\}$ would be at any of the vertexes of the box type support set, Ξ, of $\boldsymbol{\xi}$, (Ben-Tal et al., 2009). Then, the equivalent finite formulation for the distributionally robust recourse function (2.24), can be defined as follows:

$$\begin{aligned} g_{\text{DRO}}(\mathbf{x}) = \min_{\alpha,\gamma} \quad & \alpha + \boldsymbol{\gamma}^{\intercal}\bar{\boldsymbol{\xi}} \\ \text{s.t.:} \quad & f(\boldsymbol{x},\boldsymbol{\xi}_k) \leq \alpha + \boldsymbol{\gamma}^{\intercal}\boldsymbol{\xi}_k \qquad \forall k \in K, \end{aligned} \tag{2.29}$$

where $\boldsymbol{\xi}_k$ is defined as the extreme points vector of the uncertain parameter, $\boldsymbol{\xi}$. k is used to index all extreme points of the support set, Ξ. By computing the dual of problem (2.29) we can derive the equivalent finite primal formulation of problem (2.24), as described in problem (2.30).

$$\begin{aligned} g_{\text{DRO}}(\mathbf{x}) = \max_{p_k} \quad & \sum_{k\in K} f(\mathbf{x},\boldsymbol{\xi}_k) p_k \\ \text{s.t.:} \quad & \sum_{k\in K} p_k = 1 \\ & \sum_{k\in K} \boldsymbol{\xi}_k p_k \leq \bar{\boldsymbol{\xi}} \end{aligned} \tag{2.30}$$

In order to derive an equivalent finite scenario-based linear program for the DRO problem defined in (2.24), we replace $f(\mathbf{x},\boldsymbol{\xi}_k)$ in (2.29) with the objective function of (2.25). The decision variables, $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are jointly minimized with first-stage decisions,

$\mathbf{x}$, and second-stage decisions, $\mathbf{y}_k$, for each extreme point k, $\boldsymbol{\xi}_k$. The equivalent finite-deterministic MILP of problem (2.24) can be formulated as follows,

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y},\alpha,\boldsymbol{\gamma}} \quad & \boldsymbol{c}^\intercal \mathbf{x} + \alpha + \bar{\boldsymbol{\xi}} \\ \text{s.t.:} \quad & \mathbf{x} \in \mathcal{X} \\ & \boldsymbol{h}^\intercal \mathbf{y}_k \leq \alpha + \boldsymbol{\gamma}^\intercal \boldsymbol{\xi}_k \qquad \forall k \in K \\ & \boldsymbol{W}\mathbf{y}_k \geq \boldsymbol{B}\boldsymbol{\xi}_k + \boldsymbol{G}\mathbf{x} \qquad \forall k \in K, \end{aligned} \tag{2.31}$$

which is an equivalent scenario-based approach, considering only the subset of scenarios in Ξ, which are candidates for a positive-probability mass in the worst-case distribution.

Correspondence between two-stage RO and DRO approaches

The DRO reformulation in the set of equations (2.31) can be used to derive an equivalent finite formulation of the two-stage robust model, (2.21), presented in subsection 2.3.1. The equivalent robust optimization approach can be achieved by excluding the moment information, $\bar{\boldsymbol{\xi}}$, from the objective function and related constraints. In this case, the operational model's solution will be the worst-case scenario in the support set Ξ, as we present in the set of equations (2.32).

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y},\alpha} \quad & \boldsymbol{c}^\intercal \mathbf{x} + \alpha \\ \text{s.t.:} \quad & \mathbf{x} \in \mathcal{X} \\ & \boldsymbol{h}^\intercal \mathbf{y}_k \leq \alpha \qquad \forall k \in K \\ & \boldsymbol{W}\mathbf{y}_k \geq \boldsymbol{B}\boldsymbol{\xi}_k + \boldsymbol{G}\mathbf{x} \qquad \forall k \in K \end{aligned} \tag{2.32}$$

2.5. Summary and concluding remarks

This chapter provides a brief overview of the decision-making theory under uncertainty. We describe the fundamentals of stochastic programming, robust optimization, and distributionally robust optimization. Such optimization methods under uncertainty are useful to solve real-life problems in which unknown parameters can be characterized as random variables. Several versions of the approaches were presented as well as a brief review of their up-to-date solution methods in the literature.

Figure 2.1 shows in compact form, an overview of the optimization models described in this chapter, where $f(\mathbf{x}, \boldsymbol{\xi}) = \min_{\mathbf{y}} \boldsymbol{h}_{\boldsymbol{\xi}}^{\intercal}\mathbf{y}$, s.t.: $\boldsymbol{B}_{\boldsymbol{\xi}}\mathbf{y} \geq d_{\boldsymbol{\xi}} - \boldsymbol{W}_{\boldsymbol{\xi}}\mathbf{x}$. Note that we only include the modeling decision frameworks for two-stage or adjustable programs.

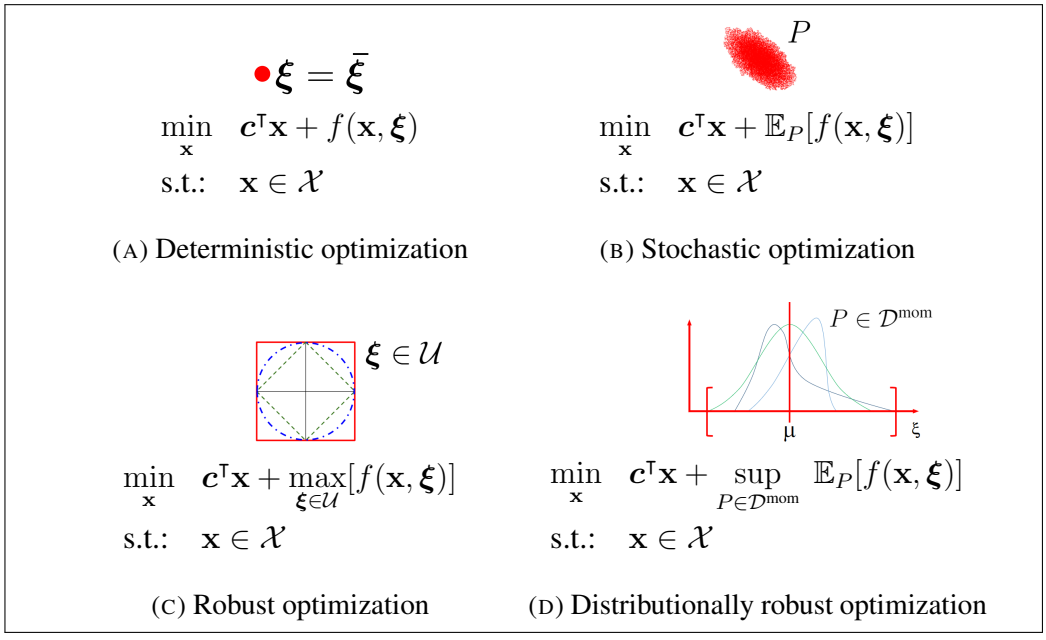


FIGURE 2.1. Comparison of the different optimization approaches under uncertainty.

The deterministic model, Figure 2.1a, assumes the uncertainty, $\boldsymbol{\xi}$, as a singleton (point estimates), $\bar{\boldsymbol{\xi}}$, that can be defined as the expected value of the uncertain parameter obtained from historical data or estimations from expert's opinion. The main disadvantage of this

approach is that small changes in such mean values may result in infeasible solutions. The two-stage stochastic model, 2.1b, minimizes the total cost under the expected cost of the operational function, $f(\mathbf{x}, \boldsymbol{\xi})$, represented by a finite set of scenarios with a known probability distribution, P of the continuous random vector $\boldsymbol{\xi}$. The main drawback of TSO is the computational load of computing, which increases with the number of scenarios, and it is even more difficult to solve for problems incorporating integer variables; scenario reduction techniques may be needed. In addition, since the uncertain parameters' distribution is inferred from data, accurate characterization of uncertainty must be guaranteed.

The adjustable robust optimization model, 2.1c, minimizes the total cost considering the worst-case realization of the random parameter, $\boldsymbol{\xi}$, within the uncertainty set, $\mathcal{U}$. The uncertainty can be characterized according to different set structures, each of them indicating a different level of protection against uncertainty, including box uncertainty (red line figure), ellipsoidal uncertainty (blue dash line figure), and polyhedral uncertainty (black line figure). Here, no distributional information is required, ensuring a trade-off between accuracy and tractability. The distributional robust optimization model, 2.1d, is a generalization of the stochastic and robust models, that computes the worst-case expected cost over the probability distributions within the ambiguity set. The ambiguity set, $\mathcal{D}^{\text{mom}}$, in Figure 2.1d represents the family of distributions of $\boldsymbol{\xi}$, in which information about, the mean, support, and variance (for moment-based ambiguity sets) can be estimated from historical data.

The optimization methodologies under uncertainty presented in this chapter are applied in the subsequent chapters of this thesis. The TSO approach is considered in Chapter 3, and solved employing a SAA method. In Chapter 5, the DRO approach under moment information is employed by considering a scenario-based solution methodology.

The performance of the DRO approach is then benchmarked with two-stage RO and TSO approaches.

Chapter 3. MULTI-OBJECTIVE ADMISSION PLANNING PROBLEM: A TWO-STAGE STOCHASTIC APPROACH

The public healthcare system is under constant pressure from the government to achieve quotas of patient's admission. Besides, hospital managers must guarantee cost-effective care with limited funding and constraints on the availability of resources. However, the uncertainty in the patient's arrival and the availability of resources makes the patient's allocation challenging to manage. Bed shortage is one of the main drawbacks of the allocation process, which can increase patient delays, transfers, unnecessary waiting times, and mortality. Thus, the admission planning process should anticipate such effects caused by the uncertainty, in order to reduce the impact of two conflicting objectives, service quality, and hospital performance.

This chapter studies intertemporal decisions in the admission planning problem through a two-stage stochastic approach. We tackle the APP at the tactical-operational level to study the trade-off between bed utilization and cost of service. We propose a bi-objective stochastic optimization model taking into account demand and capacity availability uncertainty. Real data from the surgery and medical care units of a Chilean public hospital illustrate the approach and validate the model.

The content of this chapter is based on a paper published in the *Health Care Management Science* (Batista, Vera, & Pozo, 2020).

This chapter is organized as follows. Section 3.1 provides an overview of the problem of tactical admission planning in healthcare. Section 3.2 presents the proposed framework formulation to solve the admission planning problem. Section 3.3 provides the modeling of the admission planning problem and discusses the problem setting and case of study. In Section 3.4, we present a computational study and thoughtful analysis of the results. Finally, Section 3.5 summarizes and concludes the chapter.

Notation

The mathematical symbols used throughout this chapter are described below:

Indexes

i	Patients groups.
k	Scenarios.
t	Time period.

Sets

P	Set of patients groups.
T	Set of time periods t.
$\mathcal{X}$	Set of feasible admission plans.
$\mathcal{Y}$	Set of feasible allocation plans.
Ξ	Set of scenarios.

Parameters

c_t	Internal capacity availability in period t.
d_{pt}	Demand of patients group p in the period t.
e_t	External capacity availability in period t.
n_t	Temporary capacity availability in period t.
q_p	Maximum number of patients type p in queue.
u_t	Target utilization in period t.
w_p^{r}	Cost of reserve capacity by patient group type p.
w_p^{i}	Cost of internal assignment by patient group type p.
w_p^{e}	Cost of external assignment by patient group type p.
w_p^{t}	Cost of temporary assignment by patient group type p.

w_p^{u} Cost of unmet demand by patient group type p.

Variables

H_{pt} Number of unmet demand by patient group p in period t.

K_{pt} Number of beds reserved to patient group p in period t.

X_{pt} Number of internal assignments by patient group p in period t.

Y_{pt} Number of external assignments by patient group p in period t.

Z_{pt} Number of temporary assignments by patient group p in period t.

UU_t Under-utilization of resource in the period t.

OU_t Over-utilization of resource in the period t.

3.1. Introduction

Public hospitals face political pressure to improve its processes due to limited resources. Also, the aged population and new technological advances push the healthcare system to be more efficient. The main issues are waiting lists and waiting times, that affect the timely access and the quality of service offered to patients. Additionally, the heterogeneity of patients and their resource requirements cause a high variability in the decision process. Under this complex framework, decision-makers should develop robust admission plans to achieve efficient solutions without compromising performance targets.

The APP at the tactical level consists of making decisions about the mix of patients admitted, considering resource requirements. In particular, the inpatient beds, which are scarce resources, are among the most critical assets in the admission process. The beds are considered the fundamental measure of capacity in hospitals because they affect the expenditure, quality of care, and patient access (Green, 2005). Within the admission planning, two major patient categories are distinguished: elective and emergency patients. Elective patients can be planned while emergency patients need to be admitted in a short

time. The planning of elective patients is more studied since it can be planned in contrast to urgent or emergency patients (Hulshof et al., 2013; J. Vissers et al., 2005). As both patients share the same resources, an efficient admission policy at the tactical level should consider slack capacities to handle the uncertainty caused by the arrival of unscheduled patients (Gemmel & Van Dierdonck, 1999). A typical tactical policy in hospitals to perform the admission plan is to allocate patients considering a fixed target of bed utilization, e.g., as 85% (Green, 2002), that is strategically defined. The accomplishment of the bed utilization target is crucial because it determines the efficiency of the admission plan. Therefore, the evaluation of resource utilization deviation from a predefined target helps to assess hospital performance. Another admission policy is to use waiting lists to buffer capacity for elective patients. However, these policies may result in an over or underestimation of beds, causing delays and rejections if the uncertainty at the operational level is not acknowledged.

Another issue is that most hospitals' allocation policies prioritize resource performance rather than the service criteria (J. M. Vissers et al., 2007). Hospitals aim to use their available resources to the maximum, i.e., treating as many patients as possible. In contrast, patients pursue timely access and better quality of care. The service level is related to the ability to allocate patients according to their needs at a specific cost, while the hospital performance concerns the optimal use of resources according to a target strategically determined. The decision framework mentioned above evidences an interesting *trade-off* between two conflicting objectives resource utilization and cost of service.

The admission planning problem considering bed capacities have been studied in the literature. The studies are focused on admission and resource allocation to improve patient access and resource utilization. The reader is referred to Kusters and Groot (1996); Green

(2005), and Samudra et al. (2016) for a more detailed survey of related works. We focus on studies of admission planning at the tactical and operational levels.

Table 3.1 compares the contributions in the APP with our proposed approach. The papers listed are taken from Tables 1.1 and 1.2 presented in Chapter 1, refining those that consider beds as a resource of allocation. We have included the main characteristics of the admission problem at the tactical-operational level. Column 1 indicates the paper information. Column 2 details whether the research considers intertemporal decisions. Column 3 details the use of capacity reserve variables in the modeling approach. Column 4 includes information about the assumption of demand or LoS uncertainty. Column 5 details the incorporation of flexible allocation rules. Column 6 shows if the contribution considers real data. Column 7 and 8 show whether the research considers the service concept or utilization concept in the modeling approach, respectively. Finally, column 9 indicates whether the research considers a bi-objective approach. The black or white dot indicates whether a feature is considered or not, respectively.

Several gaps can be observed in Table 3.1. The use of reserve capacities at the tactical level to prevent patient diversion and rejection is rarely applied. The studies Seung-Chul and Ira (2000); Adan and Vissers (2002); Adan et al. (2011); Barz and Rajaram (2015) and Samiedaluie et al. (2017) are the only exceptions, but the problem is solved considering a single level approach. Thus, no consistency in decision-making is guaranteed. Flexible allocation rules, such as diverting patients to external allocation and temporal assignment, are considered in some contributions, see, e.g., Adan and Vissers (2002); Demeester et al. (2010); Ceschia and Schaerf (2011); Range et al. (2014); Turhan and Bilgen (2017); Guido et al. (2018); Ceschia and Schaerf (2012, 2016); Vancroonenburg et al. (2016). However, such studies are instances of a single level decision framework considering an individual objective.

TABLE 3.1. Comparison of the proposed approach versus current literature on the admission planning allocation problem.

Research	Intertemporal decisions	Capacity reserves variables	Stochastic demand/LoS	Flexible allocation	Real data	Service concept	Resource utilization concept	Bi-objective
Demeester et al. (2010); Ceschia and Schaerf (2011)	○	○	○	●	○	○	●	○
Range et al. (2014); Turhan and Bilgen (2017); Guido et al. (2018)	○	○	●	●	○	●	○	○
Hulshof et al. (2013, 2016)	○	●	●	○	●	●	○	○
Seung-Chul and Ira (2000).	○	○	●	○	○	●	○	○
Green and Nguyen (2001).	○	●	○	○	●	○	●	○
Adan and Vissers (2002).	○	○	●	○	●	●	●	○
Harper and Shahani (2002).	○	○	●	●	○	●	○	○
Utley et al. (2003).	○	○	●	○	●	○	●	○
Adan et al. (2009).	○	○	●	○	●	●	●	○
Mazier et al. (2010).	○	○	●	○	●	●	○	○
Adan et al. (2011)	○	●	●	○	●	●	●	●
Bekker and Koeleman (2011).	○	○	●	○	○	○	●	○
Conforti et al. (2011).	○	○	○	○	●	●	○	○
Helm et al. (2011).	○	○	●	○	●	○	●	○
Bachouch et al. (2012).	○	○	○	○	○	●	○	○
Ceschia and Schaerf (2012).	○	○	●	●	○	○	●	○
Zhang et al. (2012).	○	○	●	○	●	●	○	○
Barz and Rajaram (2015).	○	●	●	○	●	○	●	○
Meng et al. (2015).	●	○	●	○	●	○	●	○
Ceschia and Schaerf (2016).	○	○	●	●	○	○	●	○
Vancroonenburg et al. (2016).	○	○	●	●	○	○	●	○
Samiedaluie et al. (2017).	○	●	●	○	●	●	○	○
Li et al. (2018).	○	○	●	○	●	●	○	○
Liu et al. (2019).	○	○	●	○	●	●	●	○
Our model	●	●	●	●	●	●	●	●

Most papers have focused on resource utilization metric rather than the service concept, see, e.g., Demeester et al. (2010); Ceschia and Schaerf (2011); Range et al. (2014); Turhan and Bilgen (2017); Guido et al. (2018); Green and Nguyen (2001); Adan and Vissers (2002); Harper and Shahani (2002); Adan et al. (2009, 2011); Bekker and Koeleman (2011); Helm et al. (2011); Ceschia and Schaerf (2012); Barz and Rajaram (2015); Ceschia and Schaerf (2016); Vancroonenburg et al. (2016); Meng et al. (2015). Only one paper (Liu et al., 2019) evaluate both concepts in the allocation problem, but not within a bi-objective framework. Besides, nearly all papers listed consider a source of uncertainty (i.e., demand and LoS) assuming knowledge of the probability distribution of the uncertain parameter. Some studies employ real data to fit the distribution, as we approached.

Many contributions focus on solving a deterministic approach of the APP, see, e.g., Demeester et al. (2010); Ceschia and Schaerf (2012); Range et al. (2014); Turhan and

Bilgen (2017); Guido et al. (2018); Adan and Vissers (2002); Conforti et al. (2011); Bachouch et al. (2012). Their main objective is to study computational performance and other characteristics, as we detailed in Table 1.1 of Section 1.5.

We observe that only one contribution focuses on solving the APP considering intertemporal decisions. Adan et al. (2011) developed a MIP divided into two stages; the first stage concerns a tactical plan to reserve beds for urgent patients. The second stage develops operational strategies to manage the flow of elective and urgent patients. The results show the relationship between patient satisfaction and resource utilization. Although this approach shares some similarities with our proposal, several differences can be noticed. Firstly, the model is used to schedule surgical patients considering resource constraints such as beds. Secondly, the authors assume that all resources are dedicated to a cardiothoracic service. Our model instead considers a centralized system in which beds are shared for several services in the hospital. This feature may allow for more realistic results considering that inpatient beds are critical resources in the hospital. Also, the solution method employed in Adan et al. (2011) is a MIP in which results are obtained through simulations applying flexibility rules between elective and urgent patients. In contrast, we propose a TSO in which the scenarios are generated through SAA.

As we indicated in Table 1.1, methods such as queuing theory (Bekker & Koeleman, 2011; Green & Nguyen, 2001; Utley, Jit, & Gallivan, 2008) and MDP (Helm et al., 2011; Li et al., 2018; Liu et al., 2019) are the most considered to solve the admission problem at the tactical-operational levels. The main disadvantage of such methods is that they assume a steady-state system and that it becomes complicated to apply it to the entire patient flow (He et al., 2019). Besides, they are mostly applied to a single level of decision. Dynamic Programming methods have also been considered (Barz & Rajaram, 2015; Hulshof et al.,

2016; Range et al., 2014; Samiedaluie et al., 2017), which could permit readjust the plan by including new information; however, the curse of dimensionality is the main drawback.

In contrast to the methodologies mentioned above, stochastic programming is useful to make decisions under uncertainty when the probabilities distribution of random parameters can be estimated (Birge & Louveaux, 2011). In particular, a TSO approach allows evaluating intertemporal decisions in the admission planning problem.

Overall, the lack of visibility in bed requirements complicates the admission process at the tactical level. Admission decisions are usually made without considering the future realization of the system. As a consequence, rejections and long waiting times are faced by patients with high priority. To overcome this problem, we propose a bi-objective stochastic approach for real-life applications to study intertemporal decisions at the tactical and operational levels of planning.

The contributions of this chapter are twofold. Firstly, we develop a TSO model to address the APP for optimal patient allocation on beds at the tactical and operational levels, considering demand and capacity availability uncertainty. The approach allows evaluating the APP from both perspectives, hierarchical structure, and uncertain nature of the problem. At the tactical level, the goal is to decide the beds that should be dedicated to different patient groups. These decisions are then constrained at the operational level, where the flow of patients and bed availability are stochastic.

Secondly, we incorporate a bi-objective approach to evaluating the trade-off between two conflicting objectives in the APP: resource utilization deviation and the cost of service. To the best of our knowledge, this study is the first effort in the literature to explore the APP as a two-stage stochastic model in multi-objective fashion. The model accounts for a balance of service level considering hospital and patient perspectives in the allocation process. We include flexible options for allocation, such as diverting patients to another

hospital and temporary assignments. Also, unlike other studies, we consider bed allocation decisions for the entire hospital instead of a single unit. Finally, the proposed approach is validated with real practices on the APP for a Chilean public hospital, a center of reference in the public system.

3.2. Framework formulation and solution methodology

In this section, we propose a bi-objective two-stage stochastic optimization model for the APP, under patient demand and capacity availability uncertainties. For details about the theory of two-stage stochastic optimization, the reader is referred to Chapter 2, Section 2.2.

In our approach, decisions are made in two stages. The first-stage variables are represented by the vector, $\mathbf{x}$, and are decisions known as *"here-and-now"*. They are made before the realization of uncertainty. The first-stage decision variables are related to decisions about reserve capacity of internal beds. The second-stage decision variable, $\boldsymbol{y}(\boldsymbol{x})$, known as *"wait and see"* are taken after the realization of uncertainty. Those variables concern decisions about patient assignment and resource utilization deviation. We will assume that uncertainty is present in some of the parameters of the problem, which we will denote by $\boldsymbol{\xi}$. Hence, second-stage variables depend on first-stage variables and the realization of the uncertain parameters, and we will make this more explicit by denoting second-stage variables as $\mathbf{y}(\mathbf{x}, \boldsymbol{\xi})$. We will assume that the probability distribution of the uncertain parameters, $\boldsymbol{\xi}$, is perfectly known in our problem and driven from historical data.

The general form of the model is expressed in Equation (3.1), where θ_1 and θ_2 are the objectives to be minimized. θ_1 refers to the resource utilization deviation and θ_2 to the cost of service.

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})} \quad & \Big\{\theta_1\big(\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})\big),\theta_2\big(\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})\big)\Big\} \\ \text{s.t.:} \quad & \mathbf{x}\in\mathcal{X} \\ & \mathbf{y}(\mathbf{x},\boldsymbol{\xi})\in\mathcal{Y}(\mathbf{x},\boldsymbol{\xi}) \end{aligned} \tag{3.1}$$

In order to deal with the bi-objective problem, we construct the Pareto efficient curve. This curve collects the set of all solutions that are efficient for both objectives, which means it is not possible to improve one objective without worsening the other (Tamiz et al., 1999). For obtaining the Pareto frontier, we employed the weighted-sum method (Deb, 2014). We generated a single objective function composed of the weighted-sum of both objectives. The introduced weights, λ, are changed iteratively to build every point of the Pareto frontier known as Pareto-optimal solutions.

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})} \quad & (1-\lambda)\Big[\theta_1\big(\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})\big)\Big]+\lambda\Big[\theta_2\big(\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})\big)\Big] \\ \text{s.t.:} \quad & \mathbf{x}\in\mathcal{X} \\ & \mathbf{y}(\mathbf{x},\boldsymbol{\xi})\in\mathcal{Y}(\mathbf{x},\boldsymbol{\xi}) \end{aligned} \tag{3.2}$$

Of course, θ_1 and θ_2 are themselves random variables, and to address the two-stage stochastic problem (1) we implemented a SAA approach which is widely used in TSO problems. The method creates one set of second-stage variables for every possible scenario (Birge & Louveaux, 2011). Thus, we generated a large enough, although finite, set of samples from the continuous distribution of our random parameter $\boldsymbol{\xi}$, represented by $\boldsymbol{\xi}\in\Xi$, where Ξ is the set of scenarios generated. Then, an equivalent finite-dimensional problem is defined in Equation (3.3).

$$\begin{aligned} \min_{\mathbf{x},\mathbf{y}} \quad & \frac{1}{|\Xi|}\sum_{\xi\in\Xi}\Big\{(1-\lambda)\Big[\theta_1\big(\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})\big)\Big]+\lambda\Big[\theta_2\big(\mathbf{x},\mathbf{y}(\mathbf{x},\boldsymbol{\xi})\big)\Big]\Big\} \\ \text{s.t.:} \quad & \mathbf{x}\in\mathcal{X} \\ & \mathbf{y}(\mathbf{x},\boldsymbol{\xi})\in\mathcal{Y}(\mathbf{x},\boldsymbol{\xi}), \quad \boldsymbol{\xi}\in\Xi \end{aligned} \tag{3.3}$$

Where $|\Xi|$ denotes the cardinality of Ξ; hence, the number of scenarios used form the SAA problem.

3.3. Bi-objective stochastic admission planning model

3.3.1. Context and problem setting

Our study is inspired by the public healthcare system in Chile that serves 80% of the population. Due to limited resources, public hospitals require reliable admission decisions to improve their processes and the service level offered to patients. The study is conducted in an important center of reference in the Chilean public system. The hospital under study (henceforth referred to as the "Hospital") receives nearly 100,000 requests for hospitalization yearly. Inpatient beds are the most critical resources. By 2015, the Hospital has 401 staffed beds, which are shared between medical and surgery departments. The principal issues are the long waiting times and rejections. The patient's flow is managed by the CAD, which receives admission requests from different sources inside and outside the Hospital, as shown in Figure 3.1.

Patient admission planning in the CAD is based on policies that do not consider the system's uncertain state, leading to inefficient service levels and non-fulfillment of the resource utilization targets. Daily, a decision-maker (e.g., a nurse) receives requests for admission. She should allocate those admissions considering the waiting list of patients not served in the previous days.

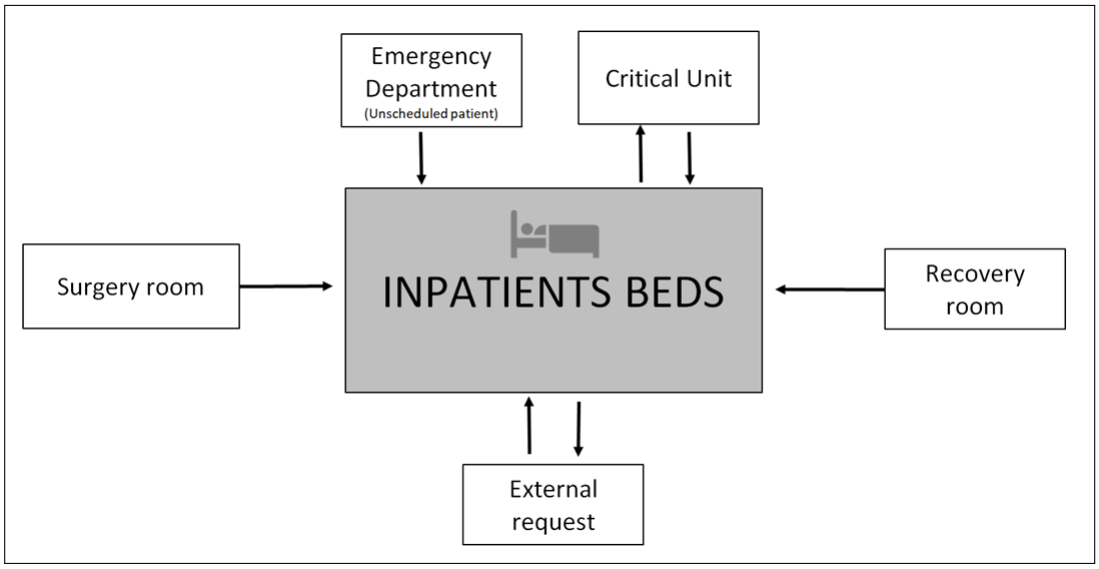


FIGURE 3.1. Patient's flow to inpatients beds in the Chilean hospital.

The current patient allocation process is shown in Figure 3.2. Specifically, the process is as follows. Firstly, the nurse receives a request for admission from different (internal and/or external) sources to allocate patients to inpatients beds. After the request is received, the nurse checks bed capacity availability. If there are not enough bed availability for all patients, they are selected according to an internal policy of patient priority. These patients are allocated to available beds. Then, the non-selected patients must wait until a bed is available or their requests are refused. Additionally, if the patient is classified as high priority, and there is not enough capacity, the patient is externalized or temporarily assigned to a stretcher or chair.

3.3.2. Admission Planning Problem formulation

In this subsection, we describe the detailed mathematical formulation of the APP model defined as a mixed-integer linear programming problem. Additionally, we present the major assumptions of the model. Demand and capacity availability are the principal sources of uncertainty. We assume both parameters are independent stochastic processes.

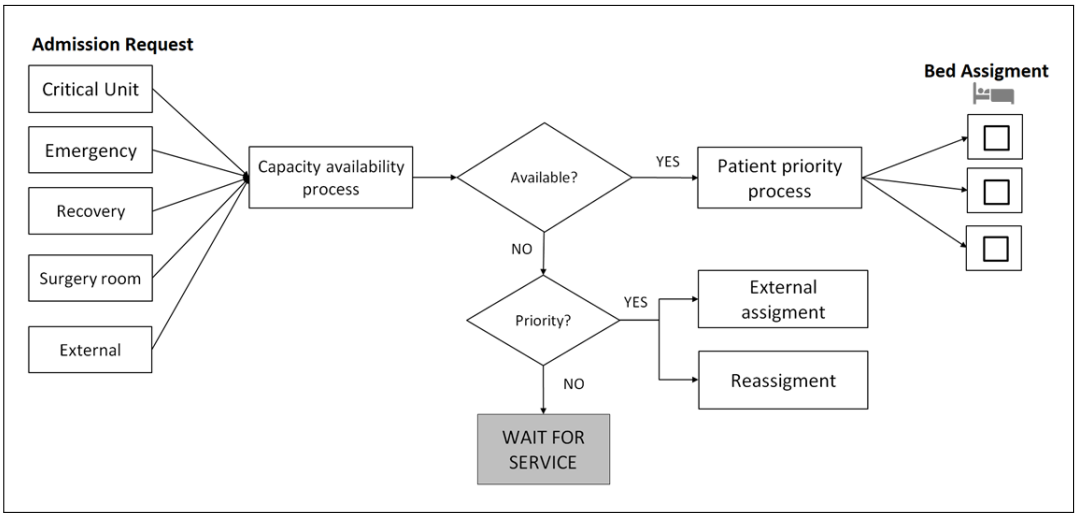


FIGURE 3.2. Patient admission planning process for a Chilean public hospital.

The demand, d_{pt}, is defined as the patients waiting to be allocated to an inpatient bed in a period. We choose the major sources of demand in the Hospital: Critical Unit (CU), Emergency Room (ER), Recovery Room (RE), Surgery Room (SU) and Network (NE). The capacity is described as the number of available beds in each time period (days). We defined three (3) sources of capacity as: internal, c_t, temporary, n_t, and external, e_t. The internal capacity, c_t, is established as the expected value in the time horizon. We assume there is infinite external capacity e_t available in the allocation process. The assumption is well-founded, considering that there is always an availability of general beds in other hospitals, including the private sector.

We considered recourse variables to allocate the patients in the time horizon. The patients can be allocated to an internal, X_{pt}, external, Y_{pt}, or temporary, Z_{pt}, bed. In case of insufficient internal capacity, c_t, the patient can be allocated to an optional bed: temporary, n_t, or external, e_t, at a certain cost, w_p. Otherwise, the allocation is penalized as unmet demand, H_{pt}, in the objective function. We assume that the cost of allocation,

w_p, is a rank of priority in the admission process; the patients are classified according to different levels of priority regarding their source of demand.

Considering the mentioned assumptions we formulate the bi-objective stochastic APP. The objective function consist of minimizing two functions: resource utilization deviation, θ_1, and cost of service, θ_2. Where:

$$\min$$

$$\theta_1 = \sum_{t \in T} \Big[UU_t(\xi) + OU_t(\xi) \Big] + \tag{3.4}$$

$$\theta_2 = \sum_{p \in P} \sum_{t \in T} w_p^{\mathrm{r}} K_{pt} + \sum_{p \in P} \sum_{t \in T} w_p^{\mathrm{i}} X_{pt}(\xi) + \sum_{p \in p} \sum_{t \in T} w_p^{\mathrm{e}} Y_{pt}(\xi)$$
$$+ \sum_{p \in p} \sum_{t \in T} w_p^{\mathrm{t}} Z_{pt}(\xi) + \sum_{p \in p} \sum_{t \in T} w_p^{\mathrm{u}} H_{pt}(\xi). \tag{3.5}$$

The objective function, θ_1 in Equation (3.4), defines the resource utilization deviation. The variables under-utilization, UU_t, and over-utilization, OU_t, compute the expected value of the deviation in the resource utilization (second-stage decisions). To evaluate the deviation, we considered a constant target, u_t, of bed utilization.

The objective function, θ_2 in Equation (3.5), defines the cost of service. This function is measured concerning the cost of allocation weighted by w_p. The first-stage variable, K_{pt}, represents the total number of beds to be reserved for patient allocation. The decisions about how many internal beds should be reserved in the first stage of the stochastic problem will affect the resource use deviation in the second-stage. The second-stage variables are recourse variables considered as corrective actions of allocation: the assignment of patients to internal beds, X_{pt}, external beds. Y_{pt}, temporary beds, Z_{pt}, and a slack variable of the unmet demand, H_{pt}, by patient group and time period. Recall that

in the SAA problem, these objective functions are averaged over the generated scenarios $\boldsymbol{\xi} \in \Xi$.

We now specify the constraints used in the model and describe them for the corresponding scenarios of the SAA problem.

Constraint (3.6) defines the unmet demand through periods. It ensures that the demand allocation of every patient group does not exceed the total requests from each source and period.

$$\begin{aligned} & X_{pt}(\xi) + Y_{pt}(\xi) + Z_{pt}(\xi) + H_{pt}(\xi) \\ & = d_{pt}(\xi) + H_{pt-1}(\xi) \qquad\qquad \forall p \in P, t \in T, \xi \in \Xi \end{aligned} \tag{3.6}$$

The capacity constraints guarantee that the allocation of patients does not exceed the availability of internal (3.7), external (3.8), and temporary (3.9) beds in each period. Constraints (3.10) refer to the first-stage decisions in which the number of beds assigned to each group of patients, must be equal to the total availability defined at the beginning of the time horizon.

$$\sum_{p \in P} K_{pt} \leq c_t \qquad \forall t \in T \tag{3.7}$$

$$\sum_{p \in P} Y_{pt} \leq e_t \qquad \forall t \in T \tag{3.8}$$

$$\sum_{p \in P} Z_{pt}(\xi) \leq n_t(\xi) \qquad \forall t \in T, \xi \in \Xi \tag{3.9}$$

$$X_{pt}(\xi) \leq K_{pt} \qquad \forall p \in P, t \in T, \xi \in \Xi \tag{3.10}$$

Expressions (3.11) and (3.12) are the utilization constraints. They ensure the fulfillment of the bed utilization target. Constraint (3.12) guarantees that the over-utilization allocation does not exceed the capacity available in each time period.

$$\sum_{p \in P} X_{pt}(\xi) + UU_t(\xi) - OU_t(\xi) = u_t \qquad \forall t \in T, \xi \in \Xi \qquad (3.11)$$

$$u_t + OU_t(\xi) \leq c_t \qquad \forall t \in T, \xi \in \Xi \qquad (3.12)$$

Constraints (3.13) establish a bound in the queue level. We guarantee that the not allocated patients at each time period must not exceed the target, q_p, defined per each patient group.

$$H_{pt}(\xi) \leq q_p \qquad \forall p \in P, t \in T, \xi \in \Xi \qquad (3.13)$$

Lastly, the nonnegative integer variables are defined in constraints (3.14) and (3.15).

$$K_{pt}, X_{pt}(\xi), Y_{pt}(\xi), Z_{pt}(\xi), H_{pt}(\xi) \in \{0\} \cup \mathbb{Z}^+ \qquad \forall p \in P, t \in T, \xi \in \Xi \qquad (3.14)$$

$$UU_t(\xi), OU_t(\xi) \geq 0 \qquad \forall t \in T, \xi \in \Xi \qquad (3.15)$$

3.4. Numerical studies

This section presents the numerical studies and the insights obtained from applying the bi-objective stochastic approach to a Chilean public hospital. Subsection 3.4.1 describes the data used in the model, and Subsection 3.4.2 presents the results with a discussion of their managerial implications.

3.4.1. Data description

The dataset has been collected from the EHR of the Hospital for the period 2010-2016s[1]. The database includes information about patient resource consumption during his/her stay in the Hospital, demand requirements, and daily capacity availability (census). Additionally, data of demographic information, the source of demand, patient length of stay, severity illness, diagnosis, treatment, and cause of the discharge is obtained from this dataset. A discretized planning horizon of T days ($t = \{1, \ldots, T\}$) related to the seven days of the week was considered. We generated 100 iid random samples, (ξ_{pt}), for each patient group, p, and time period, t, assuming a known distribution fitted using historical data. For the demand, d_{pt} we considered a Poisson distribution with rate $E[f_{pt}] = \lambda_{pt}\ \forall p \in P, t \in T$, per each source of demand and time period. For the capacity, we employed historical data of the daily capacity availability for which we adjusted a Uniform distribution $\mathcal{U}_t\ \forall t \in T$.

To implement the Pareto frontier resolution method described in Subsection 3.3.2, the objective θ_1 is weighted as $1 - \lambda$ and the cost of service θ_2 as λ. Namely, when $\lambda = 0$, the preference is to minimize the resource utilization deviation; similarly, when $\lambda = 1$, the optimization problem seeks to minimize the cost of service. To build the Pareto frontier, we defined a set of weights λ, ranging from 0 to 1, with steps of 0.02.

The experiments were run on a personal computer powered by an Intel core I7, 2.7GHz processor with 16 GB of RAM. The model was implemented in AMPL and solved with CPLEX 12.7, and it requires less than 1 second to obtain the optimal solution, due to the implicit network structure of the problem and the fact that patient demands are integer numbers.

[1]The data about demand requests and resource capacity have been obtained from a data sheet developed in conjunction with the hospital under study. See details in Appendix A.

A. Patient priority

The priority between patient categories is denoted as, w_p, in Equation (3.5). The parameter refers to a weight of importance between patient groups rather than a cost of allocation. To determine the value for each patient group, we employed a priority scoring method intending to standardize the allocation in the admission process.

In conjunction with the Chilean hospital, we determined the priority values for each patient group. We analyzed the dataset from 2010-2015 that contains information about the Diagnosis Related Group (DRG) and Severity Illness Index (SII) of each patient group. The DRG is a mechanism used in the hospital setting for reimbursement and hospital management (Groot, 1993). The method aggregates patients according to the resources employed in the medical procedure. On the other hand, the SII ranks patients according to their illness condition and level of urgency to the system according to ordinal categories (Rosko, 1988). The ranking of severity is set as 1 (minor), 2 (moderate), and 3 (major). The DRG and SII indexes were used as a proxy of patient priority, for each patient group. In order to determine the patient group's priority, we assumed that the patient group with the highest DRG and severity index, represents the higher admission priority. Table 3.2 shows the results of the analysis related to the ranking of priority, w_p, in which, $w_{CU} > w_{ER} > w_{RE} > w_{SU} > w_{NE}$.

Additionally, we defined weights for assigning a patient to an internal, external, temporary bed, and a penalty for non-assigned patients. We considered aspects related to the quality of service, i.e., there is a higher consequence of allocating a patient to a temporary bed than an external bed. Table 3.2 shows the results of the scoring method resulting in an allocation matrix in which the values are presented in ascending order according to the weight given. The cost of reserve, w_p^{r}, is related to the priority ranking, also defined in Table 3.2.

TABLE 3.2. Weight values (w_p) and demand proportion for each source of demand and allocation place.

Patient Group	Weekly demand proportion	Weight values w_p					
		Priority Ranking	Reserve (w_p^{r})	Internal (w_p^{i})	External (w_p^{e})	Temporary (w_p^{t})	Unmet (w_p^{u})
Critical Unit (CU)	13%	1	5	5	10	7	12
Emergency (ER)	45%	2	4	4	7	5	9
Recovery (RE)	17%	3	3	4	7	6	9
Surgery (SU)	16%	4	2	3	6	5	8
Network (NE)	8%	5	1	2	3	2	4

B. Demand request

The flow of patients in the CAD is characterized by a high level of variability. The decision-maker should decide which patient to serve, considering the lack of proper information about the future arrival pattern and bed availability. We have defined the demand as the request of allocation to general beds from different sources denoted as patient groups. Table 3.2 summarize the weekly demand proportion from each source of demand request. We employed historical data to perform a Chi-square goodness-of-fit test to the daily requests for source of demand. As we show in Table 3.3, the analysis suggests that for most of the patient groups, a Poisson arrival distribution is appropriate at a significance level of $\alpha = 0.10$. Note that for the patient groups ER and RE the Poisson arrival assumption is not validated from historical data. However, it is still a common assumption employed in the literature that has been shown to be appropriate in hospital admissions (Young, 1965; Mondschein & Weintraub, 2003; Green, 2005).

TABLE 3.3. Chi-square goodness of fit analysis for the patient arrival per patient group p. $N_p = 280$, $\alpha = 0.10$

Patient Group	Adjusted empirical λ_p	χ^2	p-Value
Critical Unit(CU)	3.76	29.4	0.0002
Emergency(ER)	11.0	11.682	0.8600
Recovery (RE)	3.07	3.72	0.8109
Surgery (SU)	2.07	19.36	0.0130
Network (NE)	1.29	8.66	0.0340

C. Resource capacity

The resource capacity is defined as the number of beds available every day during the week. As we described in Section 3.3, the patients can be allocated to internal, external, or temporary beds at a certain cost. We defined the internal beds as the general beds (from now on, will be referred to as beds). The temporary and external beds are used as backup capacity in case there are not enough beds available. The temporary beds are usually chairs and stretchers. Although this capacity affects the quality of care of the patient, it is a common source of allocation in the Hospital. The external beds are the available capacity in the hospital network; public and private. Patients can be sent to an external hospital at a certain cost.

TABLE 3.4. Expected values of internal capacity and fixed target of resource utilization per day of the week.

Week-day	Internal capacity(c_t)	% Target of utilization(u_t)
Monday	11	85%
Tuesday	13	85%
Wednesday	10	85%
Thursday	9	85%
Friday	10	85%
Saturday	8	85%
Sunday	11	85%

D. Target of utilization

The target of utilization is a performance indicator established as a strategical decision in the public sector (i.e., 85%). We considered this target as a fixed performance measure in the bed use. Table 3.4 shows the data about the expected value of internal capacity and the fixed target of resource utilization for each day of the week. The measure helps to assess the performance of the Hospital regarding the use of resources. Also, it is defined

as a measure of slack for unscheduled patients. Thus, hospitals face political and internal pressures to accomplish the target.

3.4.2. Results and discussion

In this subsection, we present and discuss the results of the proposed model. In the following sections, we evaluated the performance of the admission plan and validated the model using real data provided by the hospital under study. Additionally, we developed a sensitivity analysis over the target of utilization.

The Pareto frontier in Figure 3.3 represents the optimal solutions for the optimization problem, considering a target of utilization of 85%. The vertical axis refers to the θ_1 objective; the sum of deviations in the use of beds (below and above) of the target of utilization. The horizontal axis represents the θ_2 objective; the sum of the cost of service. These costs are associated with the allocation of patients to an internal, external, or temporary beds and a penalty for unmet demand. From Pareto optimality properties, we can state that there are non-dominated solutions because all of them are important. The Pareto curve reveals the trade-off between the evaluated objectives; the value of the cost of service increases as the resource utilization deviation is into the most efficient configuration and *vice versa*.

The detailed values of the Pareto analysis are summarized in Table 3.5. The table shows the solutions for the evaluated objectives θ_1 and θ_2 for different weights (λ), and the internal reserve capacity, K_{pt}, results indicating the number of beds dedicated to each patient group. The most interesting aspect of this table is that for values of λ near 0 (i.e., the resource utilization deviation is more significant), the reserve capacity is greater than for values near to $\lambda = 1$. Also, we observe that the patient groups, Emergency, and Recovery receive a greater number of beds compared to the other groups, which are interesting results and coherent with the current practice in the Hospital. Overall, patients

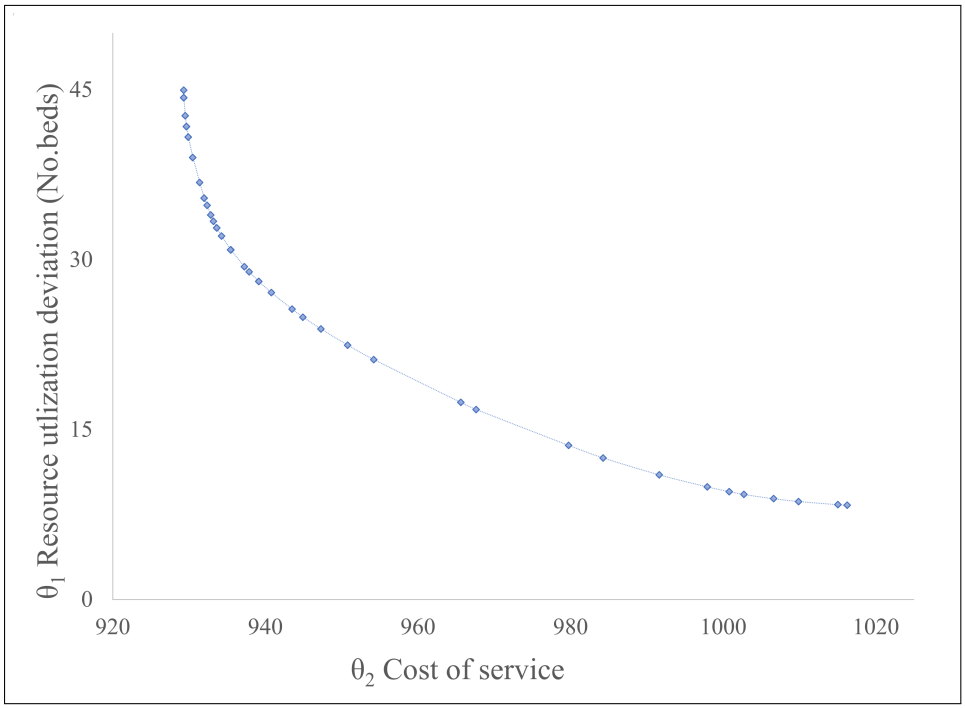


FIGURE 3.3. Trade-off between the resource utilization deviation and the cost of service, $u_t = 85\%$.

from Emergency and Recovery groups receive a higher priority concerning admission decisions. Nevertheless, the Hospital's current policy is to reserve a fixed number of beds regardless of the type of patient. Instead, the presented approach accounts for an equitable distribution of capacity, considering the variability in patient demand and resource availability.

In addition, we performed a comparison of the patient allocation considering different weights of λ, detailed in Tables 3.6 and 3.7 for the internal, external, temporary and unmet allocation, respectively. Column 1 details the weights, λ. Columns 2–11 show the result values. It concerns the percentage of patients over the total demand to be allocated in the time horizon (weekly), for a target of utilization of 85%. The allocation per patient group is summarized in Figure 3.4, which compares the proportion of patient allocation for different lambda weights.

TABLE 3.5. Values of objectives, θ_1 and θ_2, by the weight λ and reserve capacity per patient group, $u_t = 85\%$.

			Reserve capacity (K_{pt})				
λ	θ_1 Resource utilization deviation	θ_2 Cost of service	Critical Unit (CU)	Emergency (ER)	Recovery (RE)	Surgery (SU)	Network (NE)
0.00	8	1313	9	36	8	5	1
0.02	8	1016	3	32	10	8	6
0.04	8	1015	3	31	10	9	6
0.06	9	1010	3	29	10	10	7
0.08	9	1007	3	28	11	10	7
0.10	9	1003	3	25	11	11	9
0.20	14	980	4	15	12	14	10
0.30	24	947	3	3	11	14	8
0.40	28	939	1	1	10	14	7
0.50	31	935	0	0	9	12	7
0.52	32	934	0	0	8	11	7
0.54	33	934	0	0	7	11	7
0.56	33	934	0	0	7	11	7
0.58	33	934	0	0	7	11	7
0.60	33	933	0	0	7	11	6
0.70	37	931	0	0	6	11	2
0.80	42	930	0	0	2	11	0
0.90	44	929	0	0	0	10	0
1.00	54	929	0	0	0	9	0

TABLE 3.6. Patient allocation values (internal - external) in % by the weight λ, $u_t = 85\%$.

	Internal assignment (X_{pt})					External assignment (Y_{pt})				
λ	Critical Unit (CU)	Emergency (ER)	Recovery (RE)	Surgery (SU)	Network (NE)	Critical Unit (CU)	Emergency (ER)	Recovery (RE)	Surgery (SU)	Network (NE)
0.00	23.83%	31.67%	26.51%	21.99%	5.31%	66.45%	52.84%	63.08%	59.30%	53.55%
0.02	11.37%	33.60%	34.13%	38.86%	28.37%	7.35%	40.58%	60.54%	50.93%	55.27%
0.04	11.37%	32.97%	33.46%	42.11%	28.53%	7.35%	41.05%	61.11%	47.18%	56.47%
0.06	11.37%	32.01%	34.65%	41.75%	31.43%	7.35%	41.58%	60.06%	49.20%	55.56%
0.08	11.37%	30.22%	36.35%	47.18%	31.65%	7.35%	43.20%	59.18%	45.18%	56.42%
0.10	11.37%	27.98%	38.49%	47.04%	38.62%	7.35%	45.06%	56.39%	46.21%	48.57%
0.20	15.02%	17.53%	42.11%	57.34%	45.57%	6.55%	53.11%	55.69%	41.65%	46.19%
0.30	11.34%	3.97%	44.26%	65.46%	46.74%	7.39%	66.56%	54.80%	33.87%	52.39%
0.40	3.68%	1.34%	41.64%	64.59%	45.98%	8.12%	71.42%	57.42%	35.08%	53.04%
0.50	0.00%	0.00%	38.78%	60.56%	48.91%	9.04%	73.67%	60.09%	39.30%	50.22%
0.52	0.00%	0.00%	33.96%	58.48%	51.20%	9.04%	73.67%	65.39%	40.58%	48.15%
0.54	0.00%	0.00%	30.21%	58.55%	51.41%	9.04%	73.67%	68.62%	41.31%	47.83%
0.56	0.00%	0.00%	30.35%	58.35%	51.41%	9.04%	73.67%	68.90%	41.11%	47.50%
0.58	0.00%	0.00%	30.40%	58.28%	51.41%	9.04%	73.67%	68.85%	41.11%	47.61%
0.60	0.00%	0.00%	30.54%	58.08%	45.00%	9.04%	73.67%	68.85%	40.98%	54.24%
0.70	0.00%	0.00%	26.60%	58.08%	16.09%	9.04%	73.67%	72.97%	40.98%	82.72%
0.80	0.00%	0.00%	9.04%	59.09%	0.00%	9.04%	73.67%	90.26%	40.44%	98.59%
0.90	0.00%	0.00%	0.00%	54.93%	0.00%	9.04%	73.67%	99.58%	44.47%	98.15%
1.00	0.00%	0.00%	0.00%	50.50%	0.00%	9.04%	73.67%	98.50%	49.50%	99.67%

TABLE 3.7. Patient allocation values (temporary - unmet) in % by the weight λ, $u_t = 85\%$.

	Temporary assignment (Z_{pt})					Unmet demand (H_{pt})				
λ	Critical Unit (CU)	Emergency (ER)	Recovery (RE)	Surgery (SU)	Network (NE)	Critical Unit (CU)	Emergency (ER)	Recovery (RE)	Surgery (SU)	Network (NE)
0.00	0.14%	0.00%	0.00%	0.00%	0.00%	9.57%	15.49%	10.41%	18.71%	41.13%
0.02	81.23%	25.72%	3.40%	5.97%	7.00%	0.04%	0.09%	1.93%	4.24%	9.36%
0.04	81.23%	25.94%	3.68%	5.26%	5.82%	0.04%	0.04%	1.75%	5.45%	9.18%
0.06	81.23%	26.37%	3.54%	4.08%	5.11%	0.04%	0.04%	1.75%	4.97%	7.91%
0.08	81.23%	26.55%	2.95%	4.14%	4.95%	0.04%	0.03%	1.52%	3.50%	6.98%
0.10	81.23%	26.95%	3.65%	2.57%	3.45%	0.04%	0.01%	1.48%	4.18%	9.36%
0.20	78.43%	29.36%	2.20%	1.01%	1.83%	0.00%	0.00%	0.00%	0.00%	6.41%
0.30	81.26%	29.47%	0.94%	0.67%	0.87%	0.00%	0.00%	0.00%	0.00%	0.00%
0.40	88.20%	27.24%	0.94%	0.34%	0.98%	0.00%	0.00%	0.00%	0.00%	0.00%
0.50	90.96%	26.33%	1.12%	0.13%	0.87%	0.00%	0.00%	0.00%	0.00%	0.00%
0.52	90.96%	26.33%	0.66%	0.94%	0.65%	0.00%	0.00%	0.00%	0.00%	0.00%
0.54	90.96%	26.33%	1.17%	0.13%	0.76%	0.00%	0.00%	0.00%	0.00%	0.00%
0.56	90.96%	26.33%	0.75%	0.54%	1.09%	0.00%	0.00%	0.00%	0.00%	0.00%
0.58	90.96%	26.33%	0.75%	0.60%	0.98%	0.00%	0.00%	0.00%	0.00%	0.00%
0.60	90.96%	26.33%	0.61%	0.94%	0.76%	0.00%	0.00%	0.00%	0.00%	0.00%
0.70	90.96%	26.33%	0.42%	0.94%	1.20%	0.00%	0.00%	0.00%	0.00%	0.00%
0.80	90.96%	26.33%	0.70%	0.47%	1.41%	0.00%	0.00%	0.00%	0.00%	0.00%
0.90	90.96%	26.33%	0.42%	0.60%	1.85%	0.00%	0.00%	0.00%	0.00%	0.00%
1.00	90.96%	26.33%	1.50%	0.00%	0.33%	0.00%	0.00%	0.00%	0.00%	0.00%

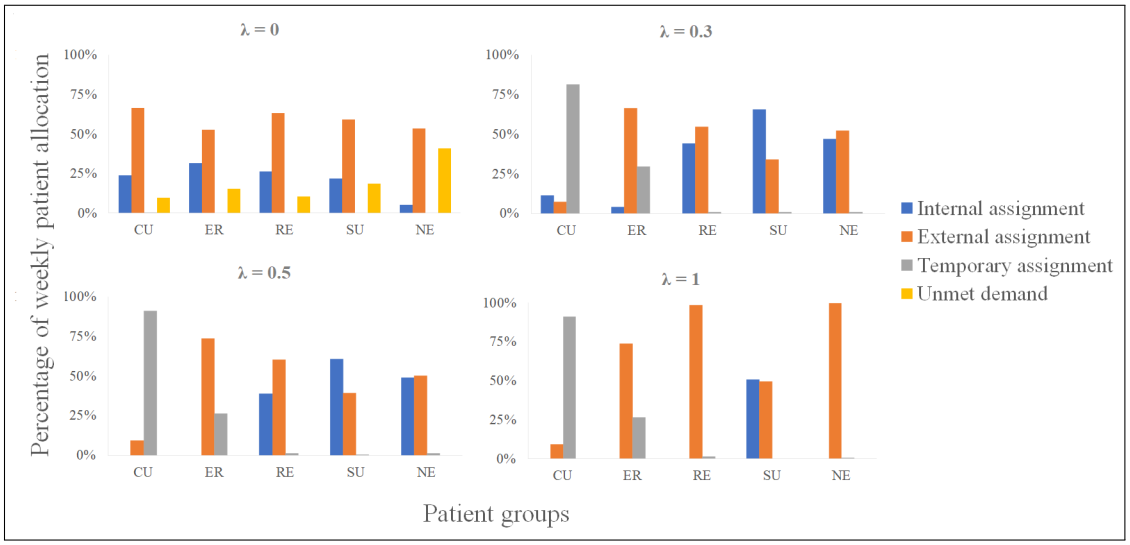


FIGURE 3.4. Comparison of the weekly patient allocation by weight, λ, and $u_t = 85\%$.

In summary, from Figure 3.4 and Tables 3.6–3.7, we observed that:

(i) Prioritize the resource utilization deviation ($\lambda \rightarrow 0$), gives a balanced assignment per patient group.

(ii) As the deviation in resource utilization is more important ($\lambda \rightarrow 0$), the percentage of unmet demand increases and the temporary assignment decreases for more priority patient groups, CU and ER.

(iii) As the cost of service ($\lambda \rightarrow 1$) is prioritized, the internal allocation is higher for less priority patient groups, RE, SU, and NE.

(iv) The unmet demand decreases as the cost of service is prioritized ($\lambda \rightarrow 1$).

Overall, the presented Pareto analysis between the cost of service and resource utilization deviation can be a convenient tool for decision-makers and practitioners in hospitals. The best setting in the admission decisions will depend on the goal expected by the Hospital. For instance, as a tactical decision, for a target of utilization of 85%, the Hospital could assume a deviation in the use of its resources of 15 beds, this would imply a cost of service nearly to 968. Alternatively, if the Hospital's objective is to accomplish the lowest deviation in the use of resources, it would entail a high cost of service. Thus, it is necessary to find a balance to satisfy both objectives for the sake of the organization.

3.4.2.1. Performance evaluation of the admission planning problem

To evaluate the admission plan performance, we analyzed the resource utilization deviation from the target in the use of beds. The over-utilization implies the bed resource is used to its maximum capacity, leaving no slack capacity for unscheduled patients. This admission policy could cause a high cost of service due to the allocation of patients on temporary or external beds or, in some cases, long waiting times. On the other hand, under-utilization means that the hospital has idle capacity. In this case, the government could justify the reduction of resources in the Hospital. The deviation function F_t in Equation (3.16), represents the sum of the internal assignment, X_{pt}, over the patients

types, p, and time period, t against the target of utilization u_t. Observe that if $F_t > 0$ it refers to over-utilization and if $F_t < 0$ to under-utilization.

$$F_t(\xi) = \sum_{p \in P} X_{pt}(\xi) - u_t \tag{3.16}$$

Figure 3.5 shows the deviation function as a continuous probability density plot. The function was estimated via kernel density estimation over the scenarios, ξ. The figure compares different weights in which $\lambda = \{0,\ 0.3,\ 0.5,\ 1\}$. The comparison stands out that as we increase the weight λ, the deviation in resource utilization increases, i.e., from [-8, 2], ($\lambda = 0$) to [-20, 0], ($\lambda = 1$). In contrast, the over-utilization is reduced by increasing the weight $\lambda = 1$, which suggests that more patients have been derived to external and temporary assignment.

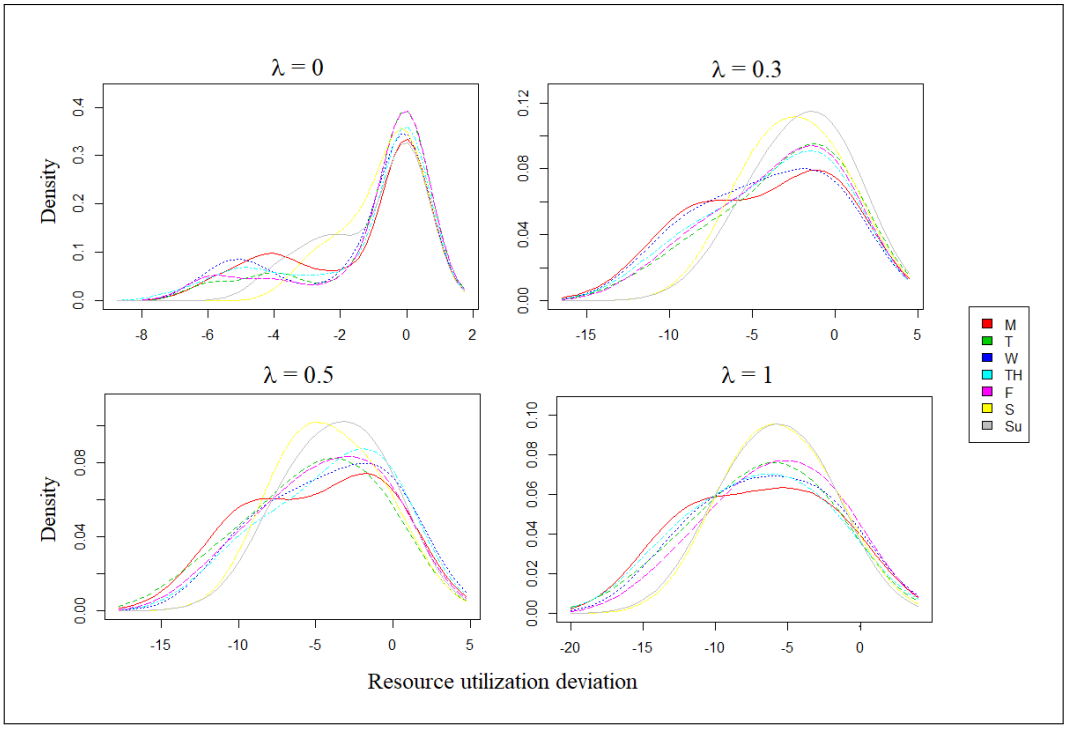


FIGURE 3.5. Comparison of the deviation function F_t, from $\lambda = 0$ (top-left) to $\lambda = 1$.

From the analysis, we can state that the resources are being used to its maximum capacity; the over-utilization is more probable than under-utilization for all cases. These results suggest that it is difficult to accomplish the expected target of utilization due to the variability, uncertainty, and internal capacity constraints for every scenario.

3.4.2.2. Validation of results

Model validation was conducted by comparing the optimal results of the proposed approach with the actual practice of the Hospital. To perform the analysis, we collected data about patient admission and bed allocation from 15 weeks, distributed among several months during the year 2017. It creates a diverse pool of weeks of historical data. The collected information regarding the cost of service and resource utilization deviation is represented in Figure 3.6 with the Pareto frontier obtained from our model. Every red square dot represents a one-week outcome for the Hospital's decisions that were made. The target of utilization is fixed to 85%.

From Figure 3.6, it can be seen that the uncertainty in the decision process challenges the Hospital to make admission decisions in which the cost of service and the deviation in the resource utilization, are in the minimum values. Interestingly, in most cases, the policy employed for the Hospital seeks to reduce to a minimum the deviation in resource utilization over the cost of service. For instance, the largest number of solutions are found in deviation levels between [0-25], which implies a high cost of service. The comparison between the Pareto frontier shows that solutions to our approach dominate the Hospital decisions in all evaluated weeks. We can state that the decisions made in the Hospital are suboptimal and can be improved with the approach proposed in this chapter.

3.4.2.3. Sensitivity analysis

The definition of a fixed target (i.e., 85%) of utilization has been shown being a suboptimal measure of capacity (Testi, Tanfani, Valente, Ansaldo, & Torre, 2008). It

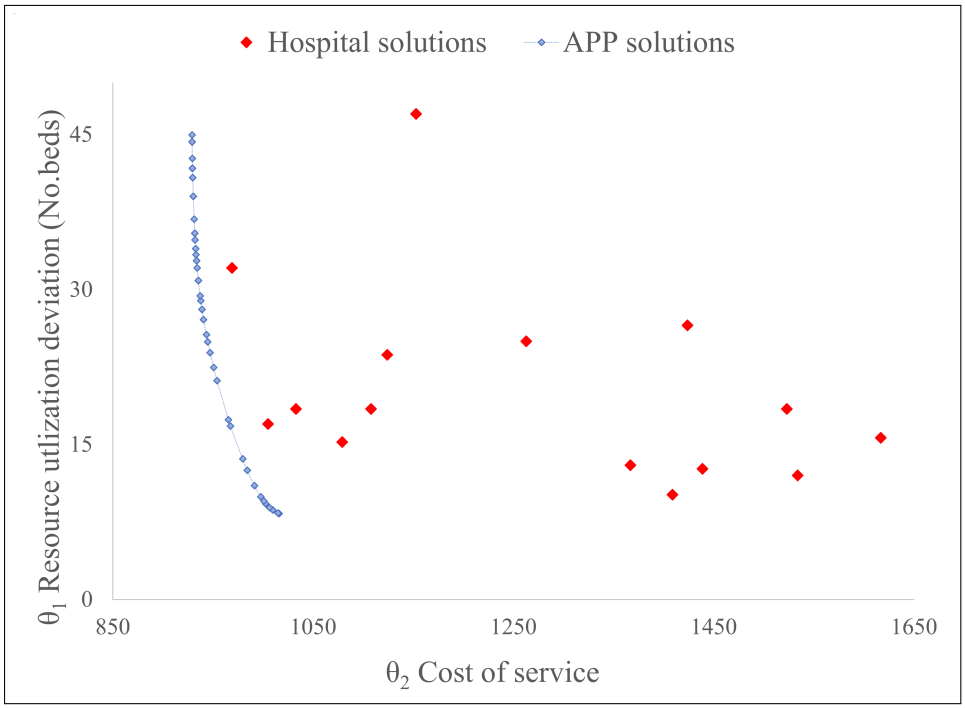


FIGURE 3.6. Validation of results of the actual practice of the Hospital, $u_t = 85\%$.

relates to a flexibility value in the decision-making process. Deciding which value is the best for any hospital configuration is not easy due to the uncertainty in the admission process.

We have developed a sensitivity analysis aiming to study how a fixed performance indicator affects the trade-off between the resource utilization deviation and the cost of service. We evaluated the Pareto frontier for different values of the target of utilization. Figure 3.7 shows the analysis for targets of utilization, $u_t = [60, 75, 85, 95, 100]\%$. The case in which the target of utilization is defined as 100% implies no slack capacity for unscheduled patients. For this case study, we observed that there is a domination of one Pareto curve over the others, depending on the target of utilization. As the target of utilization increases, the deviation in resource utilization also increases. For a fixed cost of service, the curve related to the target of 60% dominates over the others. Interestingly,

a low target suggests the best deviation levels, in contrast to the usual belief in hospital practice that establishes a greater target indicates efficiency. Therefore, we can state that a fixed target of utilization can be suboptimal if uncertainty and variability are part of the admission problem.

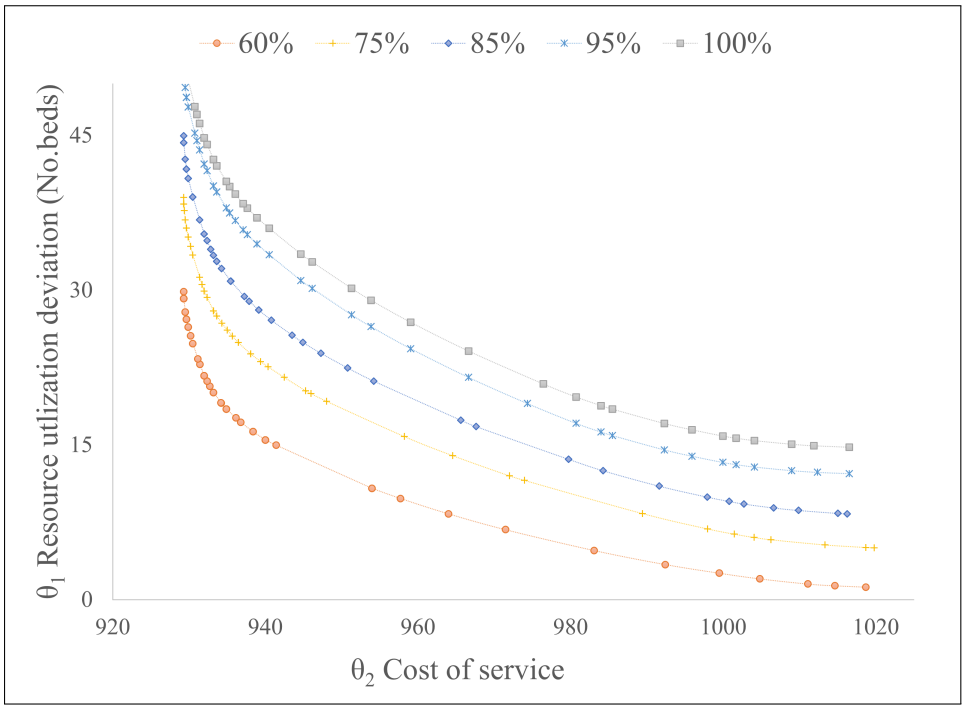


FIGURE 3.7. Sensitivity analysis of the trade-off between the resource utilization deviation and the cost of service for different values of the target of utilization, u_t.

3.5. Concluding remarks and future research directions

The hospitals pursue the lower values for two conflicting objectives: the deviation in the use of resources and the cost of service for a patient mix. One of the most critical resources to manage are the ward beds since they can cause bottlenecks in the inpatient flow. The hospital manager should define a target of utilization as a strategical decision. This measure should be accomplished at the tactical and operational levels in which bed allocation decisions are made. Nevertheless, due to the variability and uncertainty in

the arrival and the resource availability, achieving this objective is complex. Therefore, we presented a bi-objective stochastic approach to evaluating the trade-off between the mentioned objectives. A two-stage stochastic optimization model was developed in which we incorporated demand and bed availability uncertainties. For the numerical study, we considered real data from a public hospital in Chile to validate the model. The model raises awareness of the importance of making integrated and coordinated decisions in the admission process.

The results stated that the solutions to our approach outperform the actual practice in the Hospital. We found that to accomplish the target of utilization is complex due to the uncertainty at the operational level. Additionally, for cases where the capacity is limited, it is more likely to achieve over-utilization than under-utilization because the resources are being used to the maximum. Thus, the measure of a target of utilization is a strategic decision that must be defined considering the uncertainty in the lower levels rather than a fixed value over time. The admission decisions entail a process in which patients are prioritized. For the presented case of study, we found that a better balanced allocation of the patient groups can be obtained when the deviation in the use of resources is prioritized over the cost of service. Nevertheless, that policy could cause higher rates of unmet demand in the allocation process. We performed a sensitivity analysis over the target of utilization and observed that as the target is lower better solutions are obtained concerning the evaluated objectives. This result is because lower values of the target of utilization allow greater flexibility in bed allocation.

The findings of this study have important implications for future practice. The approach can support tactical to operational admission decisions at an aggregate level. The decision managers (e.g., nurses) can use the model as an optimization tool to favor the evaluated objectives instead of individual benefits. To find a balance that satisfies both

parties when several objectives conflict, it is necessary to establish a goal that indicates the breakpoint between the cost of service and resource utilization deviation.

This study has identified further research directions concerning the admission planning problem for bed allocation. The results obtained for the model were well accepted for the Hospital administration. An important practical extension is to implement the approach in the public hospital system. The model can be easily extended considering slack capacities for a network of hospitals in which the beds are shared. Also, we consider appealing to study the target of utilization as a strategic decision. We investigated this metric through a sensitivity analysis, but interesting results can be obtained if optimized. The developed approach is inspired by the hospital system. Nevertheless, it can still be implemented in any system in which admission decisions are taken in a stochastic setting and several decisions conflict.

Chapter 4. MODELING SERVICE-TIME-TYPE CONSTRAINTS FOR UNINTERRUPTED SERVICES

Problems that incorporate decisions about service duration arise in many settings, including project management (Hartmann & Briskorn, 2010), manufacturing (Mokotoff, 2001), power systems (Carrión & Arroyo, 2006), and healthcare (Gupta & Denton, 2008). In particular, appointment scheduling is a common problem consisting of determining start times, and the allocation of services for its entire duration (i.e., time allowances). If the service's duration is uncertain or difficult to estimate, the problem of determining start times and time allowances of the services becomes a challenge. Due to the random duration, the service may be completed before or after the next service's planned start time. As a consequence, waiting time, idle time, overtime, or cancellation, can occur. In addition to the uncertain duration, some services cannot be interrupted once they have started along with a determinate time window (minimum time of use). There is also the possibility that the uninterrupted service has an expiration time (maximum time of use) or a determined time window (exact time of use). Determining an efficient modeling framework for that type of problem is important to achieve efficient planning.

Typically, problems with uninterruptible service time are modeled as a summation over a rolling time window constraint employing integer variables, which difficult to consider this parameter as uncertain. In this chapter, we propose an alternative formulation in which the service time is on the right-hand side of constraints; this allows applying existing stochastic optimization methodologies easily.

The content of this chapter is partially based on a paper published in the *IEEE Transactions on Smart Grid* (Batista, Pozo, & Vera, 2020).

This chapter is organized as follows. Section 4.1 presents an overview of the appointment scheduling problem. Section 4.2 describes the appointment scheduling

process from a mathematical modeling perspective. Section 4.3 presents the current formulations in the literature and the proposed modeling framework. Finally, Section 4.4 concludes the chapter.

4.1. Introduction

The appointment scheduling problem under uncertainty shares common characteristics with several operations where customers (i.e., jobs, machines, patients) are served sequentially, service times of customers are uncertain, and time slots need to be reserved in advance for serving the customers; for instance, manufacturing, power systems, and patient allocation planning.

Research in several fields includes formulations for the appointment scheduling problem (or equivalent) in which the service cannot be interrupted once allocated (i.e., uninterrupted service). Such problems are developed by means of integer variables to describe the allocation process. In the manufacturing setting, for instance, the appointment scheduling problem shares similarities with the well-known machine scheduling problem where tasks are allocated to resources (i.e., machines) over a given time horizon. The uninterrupted appointment problem is referred to as non-preemption scheduling, and the modeling structure relies on integer programming formulations (Pinedo, 2012). In the power system literature, the unit commitment problem seeks to allocate power units considering integer minimum up-time and down-time constraints. It is classically reformulated as mixed linear constraints using a summation of binary variables over the time windows (Carrión & Arroyo, 2006; Ostrowski et al., 2012; Ozturk et al., 2004). Those settings are analogous to the healthcare system, except that in health care, it is explicitly expected that once the *customers* are assigned to a resource for the duration of their stay, they cannot be interrupted. Thus, this condition should not be relaxed.

The appointment problem in the healthcare literature is subdivided into allocating appointments to resources and simultaneously allocating and scheduling appointments to individual resources (Deng & Shen, 2016). For instance, in the outpatient setting, the goal is to determine appointment dates for patients to be assigned to a specific clinician (e.g., medical consultation) or special treatment (e.g., chemotherapy). On the contrary, in the inpatient service, in addition to determining the appointment date, the patient's allocation to resources is also decided (e.g., bed, operating room) for the duration of their stay or treatment. In particular, the problem of admitting patients to beds is referred to as the admission planning problem.

The admission planning problem is usually modeled considering deterministic stay durations of patients (He et al., 2019). In some cases, the problem includes integer variables of start and finish time of the service; the LoS, therefore, is defined as the difference of such values (Bachouch et al., 2012). Other studies (see, e.g., Demeester et al. (2010), Ceschia and Schaerf (2011), Guido et al. (2018)), in turn, assume fixed times of arrival and discharge, defined as integer parameters. Another modeling framework to determine the patient allocation includes the LoS over the summation indexes, also considering integer parameters (Conforti et al., 2011).

During the development of this research, we found a paper, (Deng et al., 2019), that performed a modeling approach similar to the one we propose in this chapter. In Deng et al. (2019), the authors developed a MILP formulation applied to the surgery scheduling problem. The modeling scheme considers two binary variables related to the patient allocation and opening of the ORs. Besides, the approach considers two continuous variables of the planned start and finish time of the service, which depends on the uncertain service time. In contrast to the formulation we propose, Deng et al. assumes continuous-time allocation to determine the start and end of the service. In

addition, since this type of allocation problem (i.e., surgery scheduling), requires to decide whether or not to use the resource, an additional binary variable (concerning our proposal) is considered. Although their formulation shares some similarities with our proposal, we intend to present a more general framework that could be applied to any problem that requires uninterrupted assignment of services in a multi-period framework.

Overall, we can state that most studies from the appointment scheduling literature (or equivalent) have considered the service time constraints as an integer parameter, while its nature is stochastic in many applications. In this chapter, we propose a new simple but effective formulation that includes constraints for problems in which a service cannot be interrupted once started, and has an uncertain time to be fulfilled.

The contribution of this chapter is twofold. First, we develop an efficient formulation that includes service time constraints for problems in which interruption is not allowed. Second, this study enhances existing admission planning models by considering a single binary variable and the service time on the right-hand side of the formulation, rather than over the indexes of a summation, how is typically modeled. This structure facilitates the implementation of the existing algorithms (e.g., dual-based methods, or Benders decomposition) that consider uncertainty, such as stochastic programming, robust optimization, and distributionally robust optimization.

4.2. The appointment scheduling process

The appointment scheduling problem plays a crucial role in any service and manufacturing industry seeking optimal scheduling. In particular, for the healthcare setting, this problem at the tactical-operational levels of planning aims to guarantee smooth patient allocation, while reducing patient waiting times and resource idle times. For

a thorough review on appointment scheduling research the reader is referred to Pinedo (2012), Cayirli and Veral (2003), and Gupta and Denton (2008).

In this section, we detail the appointment process in the healthcare setting[1] to give a better understanding of the modeling approach we propose. Figure 4.1 outlines the appointment scheduling process divided into two main parts: allocation and scheduling, which describe a six-stage process. A modeling approach may consist of some of these steps or the complete process, which varies according to the scope of decisions..

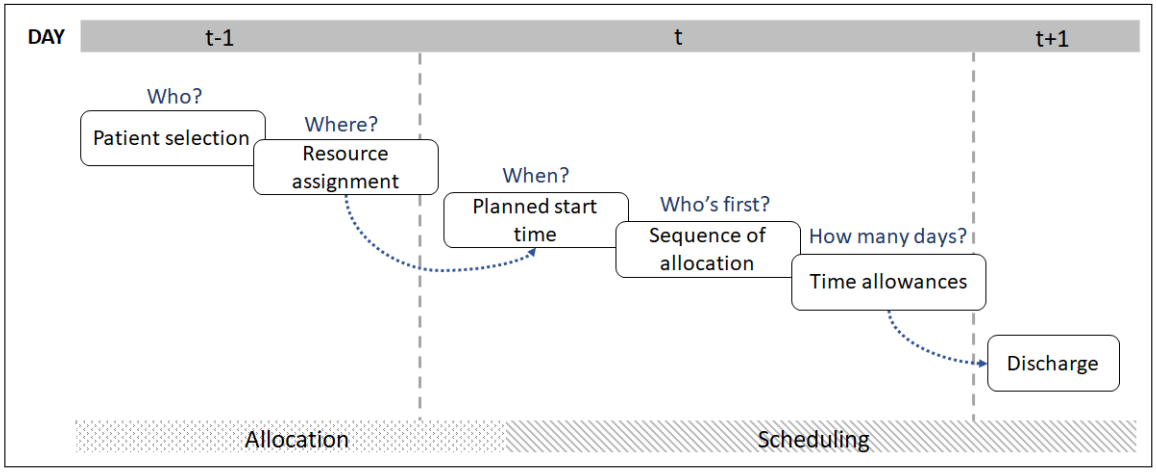


FIGURE 4.1. An illustrative representation of the appointment process.

1. **Select the service(s) to be scheduled (service mix)**. This stage determines, (*who*), which patient (or patient group) to admit or allocate according to different criteria, such as the urgency of treatment, or patient preference.
2. **Assign the service(s) to the resource**. This stage consists of the allocation of the patients previously selected to resources. The modeling framework in this stage can be defined for single or multiple resources. The resource allocation depends on the problem scope. For instance, in the outpatient setting, the resources are mostly physicians for consultation or treatment. For such cases,

[1]Here, we refer only to elective patients.

the appointment slot is usually fixed. For the inpatient setting, the resources are commonly operating rooms and beds. In the inpatient setting, the allocation to operating rooms involves the objective of minimizing waiting times or idle times due to the operating room concern fixed and variable costs of allocation (e.g., clinical staff and equipment costs). The patient allocation to beds is different since the beds are a fixed resource; thus, the objective is mostly related to improving resource utilization and service cost.

3. **Determine the planned start time(s) of the service(s)**. After selecting the patients and the resource of allocation, in this step, the planner decides the appointment date or planned start time, (*when*), of the patient within the time horizon. This decision can be constrained by resource availability.
4. **Designate (or not) the sequence to perform the service(s)**. This step is performed according to the problem scope in conjunction with the previous stage (3). For instance, in surgery planning, a set of surgeries of the same type would need to be scheduled in a fixed sequence, due to clinical requirements. However, this is not the case for admission planning to beds, in which a fixed sequence of allocation is not necessary. We remark that the optimality of the sequence order has been studied in the literature, and heuristics, such as the *order of variance*, have been developed. However, there is no proof of its optimality for more than three services (Mak et al., 2014).
5. **Determine the time allowances of the service(s)**. This stage requires to decide the length of the appointment for each patient type or group. When sequencing decisions are taken, both steps are made simultaneously. Note that for the case of uncertain durations, this step involves decisions of reducing idle times and waiting times due to the early or late competition of the service in the defined sequence.

6. **Patient discharge**. The process finalizes with the patient's discharge. This step can be planned according to the time allowances determined in the previous step (5). However, several disruptions can occur during the patient hospitalization that change the planned discharge, such as the early death of the patient, clinical conditions, or managerial reasons.

4.3. Modeling service-time-type constraints

Problems that involve service-time-type constraints are considered to be computationally hard to solve. Some algorithms haven been proposed in the literature (Begen & Queyranne, 2011), in which this type of models can be solved in polynomial-time. However, the modeling structure in terms of the constraint's type, and the dimension, make this type of problem a challenge. An enhanced modeling approach could be valuable to solve computationally challenging problems, such as stochastic and robust optimization.

Several formulations have been presented in the literature to model service-time-type constraints. The modeling approaches are based on MILP models considering mostly binary and integer variables to determine the *on-off* of the service and its allocation. A common formulation approach is to model the service time considering uptime/downtime constraints. The service time is expressed by a set of linear expressions, modeled as a summation over a rolling time windows constraint. Another approach to model the service time allocation is to define parameters of start and finish time of the service to derive the service length. Such modeling frameworks are mostly considered for deterministic approaches of service time.

In this section, we first summarize the current service time formulations in Subsection 4.3.1. Then, we introduce an alternative formulation in Subsection 4.3.2.

The MILP formulation approaches detailed in this subsection contain differentiated notation, which varies according to the context they are developed. Below we detail the common notation, and then in each formulation approach, we indicate the particular notation related to the individual modeling frameworks.

We consider a finite set of services (e.g., jobs, patients, machines), I, that are available in advance of the service date or the day of processing, to be allocated in the available resource(s). We use boldfaced symbols such as $\mathbf{x}$ to represent vectors. T is used to denote the set of finite time periods in the planning horizon, $\{1, ..., |T|\}$. Thus, there are $i \in I$ services to be scheduled during a time interval $\left[1, |T|\right]$. The service duration is represented by the parameter L.

4.3.1. Current service time modeling approaches

From the literature, we distinguish three main formulation approaches: (i) the use of uptime/downtime service constraints, (ii) the use of minimum time of service, and (iii) the use of fixed integer variables to define the service length. Below we detail each of these formulations, indicating only the service time-type constraints. For the extended formulation, the reader is referred to the related studies.

(i) **Minimum Uptime and Downtime Constraints:** This formulation is mostly applied in machine scheduling and power system scheduling settings in which the services or jobs have a pre-defined minimum uptime and downtime of operation. In the healthcare setting this type of formulation can be used to define a window of a service operation, e.g., deterministic patient stay duration. The set of linear Equations (4.1)–(4.7) in the Case labeled (i), are applied to the unit commitment problem in Ostrowski et al. (2012) in which a minimum service time of a generator has to be allocated if it is committed to generate. The set of constraints consider three binary variables, i.e., v_{jt}, y_{jt}, z_{jt}, in which j refers

to the index of generators and t the time period. Besides, integer parameters of uptime and downtime of the service are also defined. An alternative formulation for the same problem type, considering one binary variable is presented in Carrión and Arroyo (2006). The set of uptime/downtime constraints are defined as follows:

(i) Minimum uptime/downtime constraints

$$\sum_{i=t}^{t+UT_j-1} v_j(i) \geq UT_j y_{jt} \qquad \forall t = L_j + 1, ..., |T| - UT_j + 1, j \in J \quad (4.1)$$

$$\sum_{i=t}^{t+DT_j-1} (1 - v_{jt}) \geq DT_j z_{jt} \quad \forall t = F_j + 1, ..., |T| - DT_j + 1, j \in J \quad (4.2)$$

$$\sum_{i=t}^{|T|} (v_{jt}) - y_{jt}) \geq 0 \qquad \forall t = |T| - UT + 1, ...|T|, j \in J \quad (4.3)$$

$$\sum_{i=t}^{|T|} (1 - v_{jt} - z_{jt}) \geq 0 \qquad \forall t = |T| - DT + 2, ...|T|, j \in J \quad (4.4)$$

$$\sum_{t=1}^{F_j} v_{jt} = 0 \qquad \forall j \in J \quad (4.5)$$

$$\sum_{t=1}^{L_j} v_{jt} = L_j \qquad \forall j \in J \quad (4.6)$$

$$v_{jt}, y_{jt}, z_{jt} \in \{0, 1\} \qquad \forall t \in T, j \in J, \quad (4.7)$$

where UT_j and DT_j are parameters referring to the up-time and down-time, respectively. The parameters F_j and L_j indicates the time the service, j, should remain *off* and *on*, respectively, so that, $F_j = \min [|T|, D_j]$ and $L_j = \min [|T|, U_j]$. Here the parameters D_j and U_j are referred to the number of hours the unit is required to be *off* and *on* at the start of the planning period. The binary variables are v_{jt} related to the on/off status of the service; y_{jt}

referred to the start-up status; and z_{jt}, which indicates the shut-down status. Equations (4.1) and (4.3) enforce the uptime constraints. Equations (4.2) and (4.4) guarantee the downtime constraints. Finally, Equations (4.5) and (4.6) force the service to remain *on* to satisfy the uptime/downtime constraints. Note that in Equations (4.1) and (4.2), the service time is modeled as a summation over a time windows for the uptime and downtime service time, respectively. Lastly, constraint (4.7) indicates the variable's domain.

(ii) **Minimum time of service constraints:** This formulation is considered for services in which a minimum time of service is required to be performed. For instance, an integer linear formulation in the healthcare literature for the patient admission problem is found in Conforti et al. (2011). In this study, a patient should be allocated for at least a minimum stay duration for a weekly schedule. The minimum time of service is assumed to be prescribed by the physician during the baseline visit and aimed to guarantee service quality to the patient.
The set of Equations (4.8)–(4.13) in the Case labeled (ii), defines the minimum service time formulation. The formulation considers two binary variables of the patient allocation, x_t, y_t, and an integer variable, z_t, related to the remaining periods concerning the minimum service time. The set of linear constraints are defined as follows[2]:

[2]Note that we have modified the notation from the source to show a more compact formulation.

(ii) Minimum time of service

$$z_t = L^{\min} y_t \qquad \forall t = 1 \qquad (4.8)$$

$$z_t \leq L^{\min} \sum_{t=1}^{T} y_t - \sum_{t=1}^{t-1} x_t \qquad \forall t = 2, ..., |T| \qquad (4.9)$$

$$L^{\min} x_t \geq z_t \qquad \forall t \in T \qquad (4.10)$$

$$\sum_{t=1}^{t+L^{\min}-1} x_t \geq L^{\min} y_t \qquad \forall t = 1 : |T| - L^{\min} + 1 \qquad (4.11)$$

$$y_t = 0 \qquad \forall t > |T| - L^{\min} + 1 \qquad (4.12)$$

$$x_t, y_t \in \{0, 1\}, z_t \geq 0 \qquad \forall t \in T, \qquad (4.13)$$

where x_t represents the occupation variable, y_t indicates if the patient is admitted, and z_t is an integer variable of the remaining periods of allocation regarding the minimum time of service denoted as $L^{\min}$. Constraints (4.8) and (4.9) computes the remaining blocks of the admitted patients in the time horizon. The group of constraints (4.10)–(4.12) guarantee that the service time is at least equal to the days required for the admitted patient. Note that in Equation (4.11), the service time is modeled as a summation over a time windows. Lastly, constraint (4.13) defines the variable's domain.

(iii) **Fixed arrival and departure service time constraints:** This formulation type is based on a MILP model and considers integers and binary variables to define the service time allocation. The framework assumes a deterministic service time, which is defined within an availability window that corresponds to the earliest and latest hospitalization periods (see, e.g., Demeester et al. (2010) and

references therein). For instance, the authors in Bachouch et al. (2012) applied the formulation scheme in the Case labeled (iii) to a capacity planning problem in healthcare, to derive an optimal bed occupancy schedule.

The formulation considers two binary variables, x_t, y, and two integer variables, s, f, to ensures the service time allocation in a finite time horizon. Note that the set of Equations (4.14)–(4.19) is a simplified version of the source, in which we omitted the service allocation index. For the complete formulation refer to Bachouch et al. (2012).

(iii) Fixed arrival and departure service time

$$\sum_{t=W^{\mathrm{s}}}^{T} \Big(x_t(B_t - 2)(B_t + 2) \Big)/(-3) = Ly \tag{4.14}$$

$$s \leq tx_t + (-x_t + 1)M \qquad \forall t \in T \tag{4.15}$$

$$f \geq tx_t \qquad \forall t \in T \tag{4.16}$$

$$f = s + L - 1 \tag{4.17}$$

$$W^{\mathrm{s}} \leq s \leq W^{\mathrm{f}} \tag{4.18}$$

$$x_t, y \in \{0, 1\}, s, f \geq 0 \qquad \forall t \in T, \tag{4.19}$$

where x_t is defined as the allocation variable in the time horizon. The variable y is also an allocation variable not indexed in time t. The integer variables s and f refer to the beginning and ending periods of the service allocation. So that, $L = f - s + 1$. The parameters, W^{s} and W^{f}, indicate the time windows of the earliest and latest hospitalization, respectively. The parameter B_t is a matrix of resource availability, with values [-2, 1, 2] indicating the resource occupation to different types of services.

Constraint (4.14) ensures the allocation of the service for its required duration to an available resource. Constraints (4.15) and (4.16) guarantee that s and f take a value within the smallest and greatest period values of stay, respectively. Constraint (4.17) computes the ending time of allocation as a function of the service time, L. Constraint (4.18) ensures that the service should be allocated within the predefined time window $\left[W^{\mathrm{s}}, W^{\mathrm{f}}\right]$. Lastly, constraint (4.19) defines the variable's domain.

In general, we observe that current formulations in the literature for modeling the service time of uninterrupted services rely on integer variables. Besides, the service time is modeled as a summation over a rolling constraint (e.g., Cases (i) and (ii)), which may difficult to consider this parameter as uncertain. We remark that the formulation presented in the Case (iii), although the service time is on the right-hand side of the constraints, the modeling framework is complex to implement. It considers two binary and two integer variables to ensure service time allocation, and it is problem-specific.

In the next subsection, we present a simple but effective formulation in which the service time is on the right-hand side of constraints rather that over the indexes of a summation; this allows applying existing stochastic optimization methodologies easily. Besides, the proposed approach considers a single binary variable and a continuous time of allocation.

4.3.2. Proposed service time modeling approach

In this subsection, we present a modeling approach to allocation problems of uninterruptible services. The decision framework is sketched in Figure 4.2.

The allocation process is subdivided into a set of T time slots over a finite time horizon. It considers the variables, y_t that indicates the initiation of the service and the

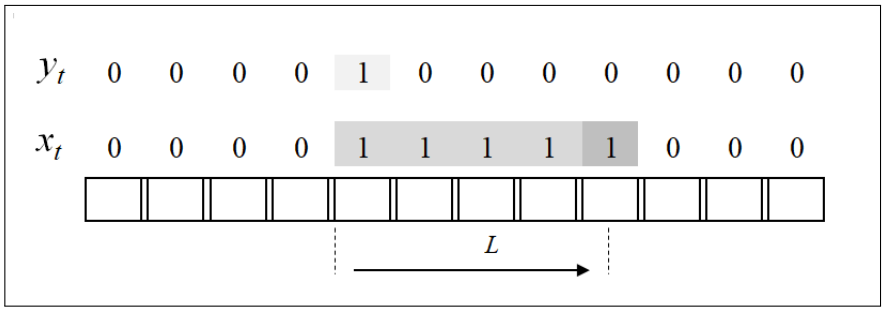


FIGURE 4.2. An illustrative representation of the allocation problem for uninterruptible service.

binary variable, x_t associated with each time slot, it takes the value 1 if we allocate this slot to fulfill the service 0 otherwise. The parameter, L, is the continuous service time of the uninterrupted service. Without loss of generality, we assume that the service always initiates at the beginning of a time slot and is assigned completely; thus, it can end at any moment and not necessarily at the end of a time slot. The service has to be allocated only once during the time horizon, and there must be enough time slots to complete the service length, L.

The proposed set of linear constraints that represent the aforementioned problem description can be formulated as follows:

Proposed alternative formulation

$$\sum_{t=1}^{T} y_t = 1 \tag{4.20}$$

$$\sum_{t=1}^{T} x_t \geq L \tag{4.21}$$

$$y_t \geq x_t - x_{t-1} \qquad \forall t = \{2, \ldots, |T|\} \tag{4.22}$$

$$y_t \geq x_t \qquad \forall t = 1 \tag{4.23}$$

$$x_t \in \{0, 1\},\ y_t \in [0, 1] \qquad \forall t \in T. \tag{4.24}$$

Constraint (4.20) indicates that the service has to be assigned once during the entire horizon. Constraint (4.21) guarantees that the service has enough time slots allocated to accommodate the service of length L. Constraint (4.22) depicts the logic between the allocation variable x_t and the activation variable y_t. We assume that for $t = 0$, $x_0 = 0$. Thus, for $t = 1$, $y_t \geq x_t$, as defined in constraint (4.23). It should be noted that constraints (4.20) and (4.21) ensure that the service cannot be interrupted, and there are enough time slots to be allocated. Lastly, constraint (4.24), defines the variables' domain. Observe that it is not required to define, y_t as binary due to Equations (4.20) and (4.22) enforce to y_t to take the value 1 for a single time slot and therefore it would be 0 for the rest of time slots.

4.4. Summary and concluding remarks

We presented an alternative formulation inspired by the appointment scheduling problem to model service-time-type constraints of uninterruptible services. In contrast to current formulations in the literature, the proposed approach considers a single binary variable and the continuous service time parameter on the right-hand side of the allocation constraint. The formulation provides relevant information on the modeling structure of allocation problems with service-time-type constraints, which can be easily adapted to yield more efficient and practical allocation policies.

The effectiveness of the approach is tested on Chapter 5 to solve the admission planning problem under uncertain patient LoS.

Chapter 5. A DISTRIBUTIONALLY ROBUST MODEL FOR THE ADMISSION PLANNING PROBLEM UNDER UNCERTAIN LENGTH OF STAY

The admission planning problem in the inpatient service aims to provide patient access and to guarantee expected levels of bed utilization. However, uncertainty in the patient's length of stay and bed availability challenge the accomplishment of that objective. Besides, there is limited information about the distribution of the uncertain parameters. Thus, several inconsistencies may arise during execution, if the plan performed at the tactical level is not robust. This chapter studies intertemporal decisions in the admission planning problem through a distributionally robust optimization approach.

We study the coordinated decisions of allocation and scheduling for the patient-to-room admission planning problem assuming heterogeneous patient types and time-varying capacity. The objective is to maximize the weighted sum of the patient's admission benefit while reducing the cost of overstay. We present a distributionally robust optimization framework that is distribution-free. The framework is robust against the infinite set of probability distribution functions that could represent the stochastic process of the patient's length of stay. To test the performance of the proposed approach, we compared it with benchmark models (i.e., deterministic, TSO, RO) employing a real data set from a public hospital in Chile. The results show that our approach outperforms the evaluated models in both reliability and computational efficiency. We provide insights to practitioners and hospital decision-makers to anticipate admission decisions while considering the randomness of the length of stay at the tactical-operational level.

The content of this chapter is based on a paper published in the *Computers and Industrial Engineering* (Batista, Pozo, & Vera, 2021).

This chapter is organized as follows. Section 5.1 outlines the problem and compares the proposed approach with up-to-date literature. Section 5.2 describes the proposed

framework of the admission planning problem, the mathematical notation, and the uncertainty modeling. Section 5.3 presents the solution methodology of the DRO approach. Section 5.4 details the numerical studies and computational experiments for the case study. Finally, Section 5.5 underlines the conclusions and findings of the study.

Notation

The mathematical symbols used throughout this chapter are detailed below:

Sets

$\mathcal{D}$	Ambiguity set of the patient's length of stay.
I	Set of patients.
K	Set of indexes of extreme point of Ξ.
R	Set of rooms related to individual beds.
S	Set of patient types.
T	Set of time periods t.
$\mathcal{X}$	Set of feasible admission plans.
Ξ	Support set for $\boldsymbol{\xi}$, defined by upper and lower bounds component-wise.
I_r	Subset of patients who belong to the rooms, r, $I_r \subseteq I$.
I_s	Subset of patients who belong to the types, s, $I_s \subseteq I$.
I_{rs}	Subset of rooms, r, and patient type, s, so that $I_r \cap I_s$.

Parameters

ξ_{is}	Random parameter of patient length of stay of patient i type s.
$\bar{\xi}_{is}$	Mean value of patient length of stay of patient i type s.
$\boldsymbol{\xi}_k$	k-th extreme point of Ξ.
θ_{is}	Penalty of overstay of patient i type s.

ζ_{is} Benefit of admission of patient i type s.

C_{tr} Available capacity in period t in room r.

D_{is} Demand of patient i type s.

O_r^{max} Maximum days of overstay to penalize per room r.

Variables

x_{rist}^{A} Binary variable denoting the initiation of the allocation of patient i type s in period t.

x_{ris}^{P} Integer variable indicating the number of admitted patients i type s.

x_{rist}^{S} Binary variable of patient allocation of patient i type s in period t.

o_{ris} Continuous variable of overstay days per room r of patient i type s.

u_{rt} Utilization of room r in period t.

α Dual variable related to the sum-one probability constraint.

γ_{is} Dual variable related to the first-moment of patient i and type s.

Functions

$f(\mathbf{x}, \boldsymbol{\xi})$ Function of the cost of admission decisions $\mathbf{x}$, and length of stay realization $\boldsymbol{\xi}$.

$g_{\mathsf{SP}^{\mathsf{APP}}}(\mathbf{x})$ Stochastic programming recourse function of the expected cost of admission decisions $\mathbf{x}$.

$g_{\mathsf{RO}^{\mathsf{APP}}}(\mathbf{x})$ Robust optimization function of the worst-case operational cost of admission decisions $\mathbf{x}$.

$g_{\mathsf{DRO}^{\mathsf{APP}}}(\mathbf{x})$ Distributionally robust recourse function of the worst-case expected operational cost of admission decisions $\mathbf{x}$.

Problems

$\mathsf{DT}^{\mathsf{APP}}$ Deterministic formulation for the APP.

$\mathsf{SP}^{\mathsf{APP}}$ Stochastic programming formulation for the APP.

RO^{APP} Robust optimization formulation for the APP.

DRO^{APP} Distributionally robust optimization formulation for the APP.

5.1. Introduction

Admission planning is a critical process in hospitals that aims to ensure timely access and efficiency in the use of resources. At the tactical-operational level for the inpatient service, the process consists of scheduling and allocating elective patients to beds for the duration of their stay and is termed Admission Planning Problem (APP). The plan is usually managed by a CAD that receives requests from different care units. Nowadays, this task is often solved manually in many hospitals, resulting in sub-optimal decisions and inefficient use of resources (J. M. Vissers et al., 2007). Therefore, there is still room for improvement in the management of admissions to derive robust decision-making plans.

One of the major complexities in the admission process is the uncertain patient's length of stay (LoS). At the tactical level, the admission decisions are taken for several categories of patients differing by priorities, lacking perfect information about the LoS. The lack of reliable information causes inconsistencies in the execution of the admission plan (i.e., operational level), such as overstay, cancellations, and fluctuation in the use of resources. The main causes of the randomness are due to individual differences in the patient's diagnosis and managerial inefficiencies that challenge the admission process. For instance, the patient overstay (i.e., prolonged stays) is caused for both the patient's clinical condition and delays during the discharge. Thus, the overstay must be considered in the planning in a way that affects the admission process to its minimum, given that it constraints the access of new patients in the admission plan and increases the hospital expenditures. Yet, although the information about the patient's LoS is unknown, decisions have to be made.

Ideally, the CAD should decide *when*, *which*, and *how many* patients to admit to the available beds, considering uncertain LoS and time-varying capacity, while reducing the overstay. The decisions about patient admission and management of overstay are related as scheduling a high number of patients may result in high levels of overstay because many patients with a short duration may have to be scheduled. On the contrary, admitting a small number of patients will tend to reduce the overstay, but lessening the level of service, defined as the number of admitted patients. A typical approach to manage the overstay is by considering a strategy of early discharge. This policy is prevalent in hospitals in which physicians have to control resources due to shortages. In this sense, the authors in Berk and Moinzadeh (1998) showed that early discharge policies (under conditions of limited capacity) could improve system accessibility without threatening care equity among the patients.

The admission planning problem has been studied in the literature. The problem has the structure of the bin-packing model, which is known to be NP-hard (Korf, 2002). The studies in the literature are subdivided into allocation and scheduling decisions according to the level of aggregation considered. Allocation decisions are performed in a more aggregated way to determine the mix of patients admitted and bed resource allocation, as presented in Chapter 3. Scheduling decisions involve a more detailed scheme, considering sequencing, and the scheduling of individual appointments. Thus, information about the patient's length of stay is required.

Table 5.1 compares the contributions in the APP with our proposed approach. The papers listed are taken from Tables 1.1 and 1.2 of Chapter 1, refining those that consider beds as a resource of allocation and the patient's length of stay to schedule appointments. We have included the main characteristics of the admission problem from the modeling structure perspective of allocation and scheduling decisions. Column 1 indicates the paper

information. Columns 2 and 3 show whether the research considers intertemporal or an integrated framework of allocation and scheduling. Columns 4–8 includes information about the patient LoS modeling. The sequence variable determines the time intervals between appointments. The inclusion of start and discharge times variables implies the determination of time allowances for the patient stay duration. We distinguish the assumption on the length of stay distribution as unknown and known distribution. Finally, column 9 shows whether the contribution considers real data. The black or white dot indicates whether a feature is considered or not, respectively.

TABLE 5.1. Comparison of the proposed approach versus existing contributions on the admission planning allocation-scheduling problem.

			Length of Stay modeling				
Research	Intertemporal decisions	Integrated A-S	Sequence (variable)	Start - discharge time (variable)	Known distribution	Unknown distribution	Real data
Mittal et al. (2014); Jiang et al. (2017); Zhang et al. (2017)	●	○	○	●	○	●	○
Harper and Shahani (2002); Li et al. (2018) Liu et al. (2019); Zhang et al. (2012)	○	○	○	○	●	○	●
Green and Nguyen (2001); Utley et al. (2003); Bekker and Koeleman (2011) Hulshof et al. (2013); Ceschia and Schaerf (2016); Vancroonenburg et al. (2016)	○	○	○	○	●	○	○
Demeester et al. (2010); Ceschia and Schaerf (2011); Bachouch et al. (2012) Range et al. (2014); Turhan and Bilgen (2017); Guido et al. (2018)	○	○	○	○	○	○	○
Min and Yih (2010b)	●	○	○	○	●	○	○
Ceschia and Schaerf (2012)	○	○	○	●	●	○	○
Mak et al. (2014)	●	○	●	●	○	●	○
Meng et al. (2015)	●	○	○	●	○	●	●
Samiedaluie et al. (2017)	○	○	○	●	●	○	●
Vancroonenburg et al. (2019)	●	●	○	○	●	○	●
Our model	●	●	●	●	●	●	●

Several observations can be made from Table 5.1. The admission planning problem was first formalized by Demeester et al. (2010) as an extension of the patient bed assignment problem. The modeling approach distinguishes between hard and soft constraints that determine the suitability of the patient assignment to a room or bed, aiming to minimize patient preference violations and transfer costs. The problem was conceived considering fixed dates of admission and discharge; thus, the LoS is assumed to be known in advance. The contributions employing such framework, see, e.g., Bachouch et al. (2012); Ceschia and Schaerf (2011); Conforti et al. (2011); Demeester et al. (2010); Guido et al. (2018); Range et al. (2014); Turhan and Bilgen (2017), have been concerned with modeling other necessary aspects, such as patient preference, age policy, room specialty,

among others. However, the assumption of fixed discharge dates is not the typical case faced by clinicians in hospitals; the patient LoS is usually uncertain and difficult to estimate due to clinical conditions. The contributions mentioned above focus on handling the dimensionality problem through heuristics and meta-heuristics based methods, which, despite presenting good performance, in terms of efficiency, do not guarantee a global optimum. An exact MILP version of the admission planning problem is proposed in Range et al. (2014) and Turhan and Bilgen (2017), but still, no uncertain parameters are considered.

The LoS uncertainty is acknowledged in some studies. Most contributions incorporate uncertainty in the APP, assuming perfect information of the uncertain parameter (see, e.g., Green and Nguyen (2001); Harper and Shahani (2002); Utley et al. (2008); Min and Yih (2010b); Bekker and Koeleman (2011); Ceschia and Schaerf (2012); Zhang et al. (2012); Hulshof et al. (2013); Ceschia and Schaerf (2016); Vancroonenburg et al. (2016); Samiedaluie et al. (2017); Li et al. (2018); Liu et al. (2019); Vancroonenburg et al. (2019)). The papers tackle the APP by employing heuristics, simulation, queue theory, or integer methods to solve the problem.

Other studies, in turn, consider that the LoS does not follow a specific distribution and assume a DRO approach, as we propose in this chapter (see, e.g., Mak et al. (2014); Mittal et al. (2014); Jiang et al. (2017); Zhang et al. (2017); Meng et al. (2015)). However, such studies are focused on scheduling decisions rather than the integration with allocation decisions, which have been shown to provide better performance (Ceschia & Schaerf, 2012). For a detailed literature review of the APP by methodology, the reader is referred to Section 1.5 of Chapter 1.

We observe in Table 5.1 that most studies do not tackle intertemporal decisions in the modeling approach. The studies are focused on a single level, mainly on operational

decisions. Besides, only one contribution (Vancroonenburg et al., 2019) tackle allocation and scheduling decisions simultaneously. The sequence variable in the scheduling problem is only considered in Mak et al. (2014). Real data is taking into account in a few studies (Harper & Shahani, 2002; Zhang et al., 2012; Meng et al., 2015; Samiedaluie et al., 2017; Li et al., 2018; Liu et al., 2019; Vancroonenburg et al., 2019).

In summary, the objective of this chapter is to study the APP to maximize patient access while reducing the cost of overstay. From the literature review, we can state that there is still a need for developing robust models for the admission planning problem by incorporating the essential features of real operations. Besides, as we stated in a more detailed review in Section 1.5, most contributions do not consider an intertemporal decision framework under limited information of the uncertain parameter of LoS. In contrast, we account for a distributionally-robust framework that considers ambiguity in the probability distribution of the patient's length of stay.

The contribution of this chapter to the issue of admission planning under uncertain length of stay is threefold. First, a new version (not yet reported in the literature) of the APP is developed at the tactical-operational level. The proposed model considers a multi-period, multi-priority, multi-specialty scheme of decisions. Besides, the bed capacity availability over time is assumed time-varying, in contrast to current contributions. Unlike most studies in the patient-to-room admission problem, we do not assume a fixed sequence of arrival. The model determines the planned start times along with the stay duration and bed assignment of patients. Such a modeling framework provide insights into the relevance of considering scheduling and allocation decisions simultaneously.

Second, to help ambiguity-averse decision-makers derive a robust plan of admission, we propose a distributionally robust optimization approach to solve the APP. Rather than assuming perfect information on the probability distribution of the patient LoS, we account

for an ambiguity-averse framework. We consider that known information is limited only to the first moment and the support set of the true probability distribution of the LoS. We solve this problem through duality theory and derive a tractable solution methodology for solving the DRO model. Thus, the infinite-dimensional DRO is then reformulated into a deterministic equivalent model.

Third, we employ real data to validate the performance of the proposed approach. The data is used to construct the ambiguity set and to generate the sample scenarios. The robustness of the approach is evaluated through an extensive computational study by comparing it with standard approaches in the literature: robust optimization, stochastic programming, and deterministic. In addition, we propose a reliability metric by benchmarking with different approaches along with the conventional cost-based metrics in out-of-sample analysis.

5.2. Problem formulation

In this section, we present the proposed framework of the APP and the uncertainty modeling of the LoS. We first describe the admission planning process and the main assumptions in Subsection 5.2.1. Then, in Subsection 5.2.2, we characterize the admission process scheme and the uncertain modeling, along with the main variables and constraints that describe the tactical-operational plan. In Subsection 5.2.3, we present the MILP deterministic formulation as an extended version of the proposed model. Finally, Subsections 5.2.4, 5.2.5 and, 5.2.6 describe the stochastic, robust and distributionally robust formulations, respectively.

5.2.1. General description and assumptions

We consider the problem of patient scheduling and allocation in admission planning. The problem consists of determining start times (i.e., date of admission) for a list of

requests that belong to a waiting list along with the room of allocation. Additionally, we derive time intervals between admissions considering random LoS. The admission decision depends on the availability of beds over time and patient priority. Once the planned start times and the time interval between patients are determined, the patient's stay duration may vary, causing disruptions in the schedule, such as overstay.

The objective of this study is to maximize the total benefit of patient admission that is composed of two parts; the reward of admission and the penalty of overstay. The first part is related to the benefit of admitting a patient according to a weighted priority. The second is a penalty cost of the prolonged patient stay that occurs due to variations in the length of stay. Thus, the overstay is used to overcome the effects of the uncertainty in the planning. In summary, the decisions to make in the APP are: (i) Which patient type to admit in the time horizon; (ii) When to assign a patient to the available resources; (iii) How many days of stay to reserve, and (iv) How many overstay days to allocate for each patient type.

5.2.1.1. Assumptions

The main assumptions considered in the formulation of the admission planning model are described as follows:

1. The patient demand, D_{is}, is known and is retrieved from the waiting list of elective patients.
2. The LoS, $\boldsymbol{\xi}$, is uncertain, i.e., a random parameter. We assume different levels of knowledge on the underlying distribution of $\boldsymbol{\xi}$: (i) full knowledge of the true distribution for the stochastic formulation of the APP, (ii) partial knowledge of the true distribution consisting of first moment (mean) and its support set for the distributionally robust formulation; and (iii) minimal information consisting of the support set of the true distribution for the robust formulation.

3. The time horizon of study is divided into periods of equal time slots, typically a day.
4. The patient LoS is defined as a continuous random value on the unit-base of the time horizon, i.e., days.
5. The admission decisions are made at the beginning of the period for the entire planning horizon. Thus, the request for admission is assumed to be at the time period, $t = 1$.
6. We assume that a patient is always assigned at the beginning of a time slot for the total period; thus, it can end at any moment and not necessarily at the end of a time slot.
7. The resources are defined as ward beds; the room type determines the distinction between them. Hence, the patients can be allocated in any of the available beds in the room.
8. The available capacity, C_{tr}, per period, t, and room, r, is known at the beginning of the plan and is considered to be time-varying in the planning horizon. The reason for assuming time-varying bed capacity is to study its influence on the scheduling and allocation process of a shared resource along with the uncertain LoS. Recently, some studies based on machine learning techniques (Rajkomar et al., 2018; Bertsimas, Pauphilet, et al., 2019), have proposed approaches to predict the hospital census and capacity availability. The developed tools may be useful as an input to our model.
9. Unmet demand is allowed due to capacity limitations. This assumption is appropriate in many hospital services, where the management of the waiting list is done periodically. No admitted patients are held on the waiting list for later scheduling.

10. Emergency patients are not considered in the scheduling. Generally, the hospitals determine a fixed quota of admission for those patients.
11. The transfer of patients is not allowed; the proposed approach accounts for the service's continuity.
12. The patient allocation priority, $\boldsymbol{\zeta}$, is defined by considering the Diagnosis Related Group (DRG) and the Severity Illness Index (SII) scores, both obtained through historical data. The DRG is a worldwide system to classify patients according to their resource consumption during the treatment (Peters-Groot, 1993), and the SII is an ordinal measure of patient severity illness (Rosko, 1988).

5.2.2. Modeling the uncertain LoS

In this subsection, we recall the modeling framework for service-time-type constraints presented in Chapter 4, and we apply it to the admission planning problem with uncertain LoS.

The off-line admission planning problem can be classified as a standard capacitated dual bin packing type problem (Vijayakumar et al., 2013), where the number of bins (beds) is given at a fixed cost. In addition, the APP has the structure of the allocation problem for uninterruptible services presented in Chapter 4. Under this scheme, once the service is allocated, it cannot be interrupted during the time window (Batista, Pozo, & Vera, 2020). Similar to what we presented in Subsection 4.3.2 of Chapter 4, Figure 5.1 illustrates the allocation scheme for a single patient, i and resource in the time horizon. A finite planning horizon of T days indexed by $t = \{1, \ldots, T\}$ is considered, subdivided into a set of t time slots of the same duration. The admission process starts at the beginning of the period, $t = 1$. From a request list, the decision-maker decides which patient, i, to admit along with the date of admission.

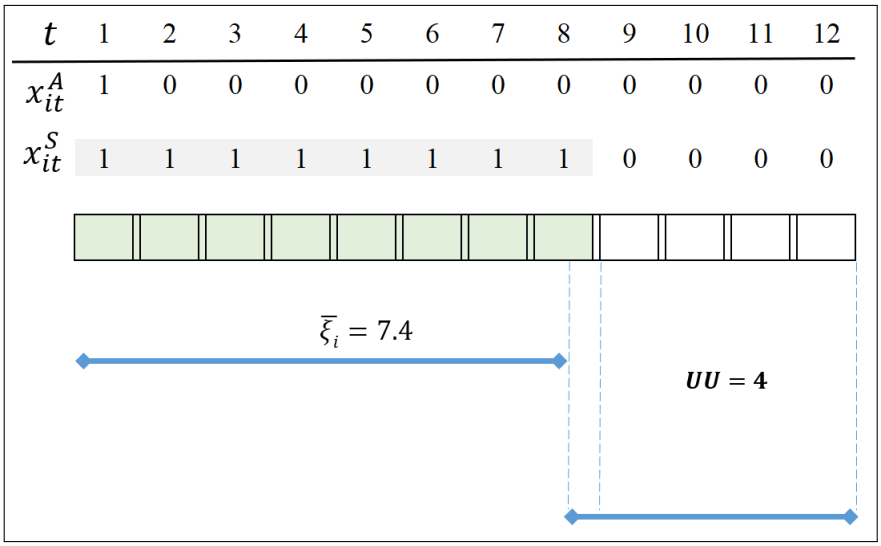


FIGURE 5.1. Representation of the proposed framework for the APP. UU measures the under-utilization.

The variable x_{it}^{A}, is associated with the initiation of the admission. The binary variable x_{it}^{S}, represents the discrete allocation of the patient stay duration and, the parameter, $\bar{\xi}_i$, is the mean value of the LoS of the patient i, not necessarily an integer value. Note that, in this subsection, ξ_i is assumed deterministic and equal to the mean value, but in the next subsection, we relax this assumption considering it a random parameter. The patient has to be allocated only once during the time horizon, and enough number of slots are required to cover the total allocation of length, $\bar{\xi}_i$. The set of linear constraints that represent the proposed approach are described in Equations (5.1)–(5.7).

$$x_{it}^{\mathrm{A}} \geq x_{it}^{\mathrm{S}} - x_{it-1}^{\mathrm{S}} \qquad \forall i \in I, t = \{2, \ldots, T\} \tag{5.1}$$

$$x_{it}^{\mathrm{A}} \geq x_{it}^{\mathrm{S}} \qquad \forall i \in I, t = 1 \tag{5.2}$$

$$x_{i}^{\mathrm{P}} = \sum_{t=1}^{T} x_{it}^{\mathrm{A}} \qquad \forall i \in I \tag{5.3}$$

$$x_{i}^{\mathrm{P}} \leq 1 \qquad \forall i \in I \tag{5.4}$$

$$\sum_{t=1}^{T} x_{it}^{\mathrm{S}} \geq x_i^{\mathrm{P}} \bar{\xi}_i \qquad \forall i \in I \qquad (5.5)$$

$$x_i^{\mathrm{P}} \geq 0 \qquad \forall i \in I \qquad (5.6)$$

$$x_{it}^{\mathrm{S}} \in \{0,1\},\ x_{it}^{\mathrm{A}} \in [0,1] \qquad \forall i \in I, t \in T \qquad (5.7)$$

Constraints (5.1) and (5.2) preserve the continuity in the allocation for the patient's entire LoS. Constraints (5.3) define the auxiliary variable, x_i^{P}; it indicates if a patient have been admitted. Constraints (5.4) indicate that a patient must be allocated to at most one room/bed in the time period. Note that with constraints (5.4), we permit unmet demand. Constraint (5.5) guarantees that there are enough time slots to allocate the duration of the stay of length, $\bar{\xi}_i$. Lastly, constraints (5.6)–(5.7) define the decision variables' domain.

As presented in Subsection 4.3.2 of Chapter 4, it is not required to define the variable, x_{it}^{A}, as binary due to Equations (5.1)–(5.4) are enforcing, x_{it}^{A}, to take the value 1 for a single time slot, and therefore it would be 0 for the rest of them. Additionally, note that the parameter, $\bar{\boldsymbol{\xi}}$, is on the right-hand side of the constraint in expression (5.5), instead of being modeled as a summation over a rolling time window. This modeling structure facilitates the implementation of dual-based algorithms, such as stochastic programming, robust optimization, and distributionally robust optimization.

5.2.2.1. Overstay modeling

In Figure 5.1 we observe that when assigning a patient for their mean length of stay (i.e., $\bar{\xi}_i = 7.4$) to the available capacity, there may be under-utilized blocks (i.e., UU = 4) because there are not enough time slots to allocate a new patient for the duration of their stay. This situation arises due to the constraint (5.5) presented in the modeling framework is very strict; it considers that the stay duration to be allocated should be at least the patient LoS. We can relax this constraint by introducing a slack variable, o_i, to capture the

variations in the LoS of the patient, i. This slack variable allows the introduction of extra flexibility in the allocation process by extending the space dimension of decisions to be taken.

Thus, we can reformulate relation (5.5) into constraint (5.8); it can be interpreted as an early discharge policy to increase patient access and to maximize bed utilization in tactical-operational planning.

$$\sum_{t=1}^{T} x_{it}^{\mathrm{S}} \geq x_i^{\mathrm{P}} \bar{\xi}_i - o_i \qquad \forall i \in I \tag{5.8}$$

5.2.3. Deterministic Admission Planning Problem formulation

Under the deterministic approach, the patient LoS is known. The MILP formulation for this problem is described in Equations (5.9)–(5.20), which is an extension of the model described in Subsection 5.2.2. We have included the indexes related to patient types, $s \in S$ and rooms, $r \in R$. In order to characterize a system in which patients of a certain type are allocated in rooms according to the related diagnose, s, we defined the subsets, $I_r \subseteq I$, $I_s \subseteq I$, and I_{rs}. I_s is referred to as the subset of patients who belong to a type, s, I_r corresponds to the subset of patients associated with a room, r, and I_{rs} is a simplified expression of the subsets I_r and I_s. The deterministic ($\mathrm{DT}^{\mathrm{APP}}$) APP is defined as follows,

$$\mathbf{DT^{APP}}: \qquad \max_{\mathbf{x}^{\mathrm{S}},\mathbf{x}^{\mathrm{A}},\mathbf{x}^{\mathrm{P}},\mathbf{o}} \sum_{r=1}^{R}\sum_{i=1}^{I_r}\sum_{i=1}^{I_s} \zeta_{is} x_{ris}^{\mathrm{P}} - \sum_{r=1}^{R}\sum_{i=1}^{I_r}\sum_{i=1}^{I_s} \theta_{is} o_{ris} \tag{5.9}$$

s.t.:

$$x_{ris}^{\mathrm{P}} \leq D_{is} \qquad \forall r \in R, i \in I_{rs}, s \in S \tag{5.10}$$

$$\sum_{i=1}^{I_{rs}} x^{\mathrm{S}}_{rist} \leq C_{tr} \qquad \forall r \in R, t \in T \qquad (5.11)$$

$$x^{\mathrm{A}}_{rist} \geq x^{\mathrm{S}}_{rist} - x^{\mathrm{S}}_{rist-1} \qquad \forall r \in R, i \in I_{rs}, s \in S, t = 2, \ldots, T \qquad (5.12)$$

$$x^{\mathrm{A}}_{rist} \geq x^{\mathrm{S}}_{rist} \qquad \forall r \in R, i \in I_{rs}, s \in S, t = 1 \qquad (5.13)$$

$$x^{\mathrm{P}}_{ris} = \sum_{t=1}^{T} x^{\mathrm{A}}_{rist} \qquad \forall r \in R, i \in I_{rs} \qquad (5.14)$$

$$x^{\mathrm{P}}_{ris} \leq 1 \qquad \forall r \in R, i \in I_{rs} \qquad (5.15)$$

$$\sum_{r=1}^{R} x^{\mathrm{S}}_{rist} \leq 1 \qquad \forall i \in I_{rs}, t \in T \qquad (5.16)$$

$$\sum_{t=1}^{T} x^{\mathrm{S}}_{rist} \geq x^{\mathrm{P}}_{ris} \bar{\xi}_{is} - o_{ris} \qquad \forall r \in R, i \in I_{rs} \qquad (5.17)$$

$$\sum_{i=1}^{I_{rs}} o_{ris} \leq O^{max}_{r} \qquad \forall r \in R \qquad (5.18)$$

$$x^{\mathrm{P}}_{ris},\ o_{ris} \geq 0 \qquad \forall r \in R, i \in I, s \in S \qquad (5.19)$$

$$x^{\mathrm{S}}_{rist} \in \{0, 1\},\ x^{\mathrm{A}}_{rist} \in [0, 1] \qquad \forall r \in R, i \in I, s \in S, t \in T \qquad (5.20)$$

The objective function, (5.9), accounts for the benefit of patient admission and the weighted cost of overstay per patient type. The weights of admission and penalty of overstay respectively, follows, $\boldsymbol{\zeta} > \boldsymbol{\theta}$.

Constraints (5.10) accounts for demand fulfilment. From Equation (5.10) we can easily calculate unmet demand. Constraints (5.11) ensure the capacity availability which varies in time, t, and room type, r. The group of Equations (5.12)–(5.15) are an extension of the constraints described in Subsection 5.2.2. Constraint (5.16) ensure that the patients will be allocated in only one room during their stay. Constraint (5.17) is an extension of the constraint (5.8); it guarantees a certain degree of flexibility in the allocation process,

by allowing days of overstay. Constraints (5.18) impose a limit of the maximum days of overstay to penalize per room, through the parameter, O_r^{max}. Finally, constraints (5.19) and (5.20) defines the variables' domain. The resulting model (5.9)–(5.20), is a MILP suitable for solving with commercial solvers.

5.2.4. Stochastic Admission Planning Problem formulation

When full knowledge about the distribution of the uncertain parameter of the patient's LoS is available, DT$^{\text{APP}}$ formulation can be extended to add such uncertainty. A two-stage stochastic optimization approach can model the APP. It divides into two sets of decision variables: first-stage and second-stage. The first-stage decisions are made before the random realizations of the patient's LoS (tactical level). The second-stage decision variables are the operational adjustments once the uncertain parameter, $\boldsymbol{\xi}$, is observed (operational level). We define the first-stage decision vector as $\mathbf{x}$, which represents the decisions, $\mathbf{x}^{\text{S}}, \mathbf{x}^{\text{A}}$, and, $\mathbf{x}^{\text{P}}$. The second-stage decisions are represented by the continuous variable of overstay, $\mathbf{o}$, renamed as the vector, $\mathbf{y}$, to follow standard nomenclature on two-stage stochastic programming.

Model (5.21) represents a compact form of the classical two-stage stochastic model with recourse. It can be defined as a simple recourse problem in which the recourse actions (i.e., overstay), are linear penalties based on the surplus of the scarce resources (Birge & Louveaux, 2011).

$$\max_{\mathbf{x}\in\mathcal{X}} \quad \boldsymbol{b}_{\mathbf{0}}^{\text{T}}\mathbf{x} + \mathbb{E}_P\left[f(\mathbf{x}, \boldsymbol{\xi})\right] \tag{5.21}$$

The first-stage admission decisions, $\mathbf{x}$, are binary variables, and the second second-stage decisions, $\mathbf{y}$, are continuous. The uncertain parameter, $\boldsymbol{\xi}$, follows a perfectly-known probability distribution. The vector, $\boldsymbol{b}_{\mathbf{0}}$, is defined as the benefit of admission. The

set $\mathcal{X}$ represents the set of constraints that affect only the first-stage decisions, $\mathbf{x}$, that corresponds to constraints (5.10)–(5.16) in the DT$^{\text{APP}}$ model.

The operational cost function, $f(\mathbf{x}, \boldsymbol{\xi})$ can be represented as in the model in relation (5.22):

$$\begin{aligned} f(\mathbf{x}, \boldsymbol{\xi}) = \max_{\mathbf{y} \geq 0} \quad & \boldsymbol{b_1}^{\intercal}\mathbf{y} \\ \text{s.t.:} \quad & \boldsymbol{W}\mathbf{y} \geq \boldsymbol{B}\boldsymbol{\xi} - \boldsymbol{G}\mathbf{x}. \end{aligned} \tag{5.22}$$

We define $\boldsymbol{b_1}$ as the penalty vector (negative) of the cost of overstay. Here, we are using the standard notation in two-stage problems, with the matrices $\boldsymbol{W}$, $\boldsymbol{B}$, and $\boldsymbol{G}$, representing the blocks of coefficients associated with the rest of the constraints of the problem. Note that both parameters, $\mathbf{x}$ and $\boldsymbol{\xi}$, of this second-stage problem, appear on the right-hand side of the linear optimization problem (5.22). Thus, we can state that $f(\mathbf{x}, \boldsymbol{\xi})$ is a convex function of the first-stage decisions of planned admissions, $\mathbf{x}$, and the length of stay realizations, $\boldsymbol{\xi}$.

We can represent the two-stage stochastic formulation (SP$^{\text{APP}}$) as (5.23)–(5.24). The SP$^{\text{APP}}$ model aims to maximize the patient admission under the expected cost of overstay.

$$\textbf{SP}^{\textbf{APP}}: \qquad \max_{\mathbf{x} \in \mathcal{X}} \quad \boldsymbol{b_0}^{\intercal}\mathbf{x} + g_{\text{SP}^{\text{APP}}}(\mathbf{x}) \tag{5.23}$$

$$\text{where:} \qquad g_{\text{SP}^{\text{APP}}}(\mathbf{x}) = \mathbb{E}_P \left[f(\mathbf{x}, \boldsymbol{\xi}) \right] \tag{5.24}$$

5.2.5. Robust Admission Planning Problem formulation

Robust optimization is an alternative framework to SP$^{\text{APP}}$, where the uncertain parameter of the patient's LoS does not rely on probabilistic information; instead, uncertainty is modeled within an uncertainty set obtained through historical data or

estimates with a confidence interval. The robust optimization (RO$^{\text{APP}}$) model is introduced in Equations (5.25)–(5.26). The support set, Ξ, is characterized by the use of estimates of the length of stay. The random vector, $\boldsymbol{\xi}$, belongs to the support, $\Xi \subset \mathbb{R}^d$.

The aim of the RO$^{\text{APP}}$ model in Equation (5.25) is to maximize the total benefit of admission within the uncertainty set, considering the worst-case realization of the random parameter, $\boldsymbol{\xi}$. The function, $g_{\text{RO}^{\text{APP}}}(\mathbf{x})$, in Equation (5.26), aims to minimize the worst-case operational cost of overstay for the admission decisions, $\mathbf{x}$.

$$\textbf{RO}^{\textbf{APP}}\text{:} \qquad \max_{\mathbf{x}\in\mathcal{X}} \quad \boldsymbol{b_0}^{\intercal}\mathbf{x} + g_{\text{RO}^{\text{APP}}}(\mathbf{x}) \tag{5.25}$$

$$\text{where:} \qquad g_{\text{RO}^{\text{APP}}}(\mathbf{x}) = \min_{\boldsymbol{\xi}\in\Xi}\Big[f(\mathbf{x},\boldsymbol{\xi})\Big] \tag{5.26}$$

5.2.6. Distributionally Robust Admission Planning Problem formulation

The distributionally robust optimization approach is a generalization of the SP$^{\text{APP}}$ and RO$^{\text{APP}}$ models, in which is considered limited distributional information of the uncertain parameter, $\boldsymbol{\xi}$. In Equation (5.27), we introduce the distributionally robust optimization (DRO$^{\text{APP}}$) model. It aims to maximize the net benefit of patient admission, considering the expected cost of overstay that is represented by an ambiguity-averse expectation measure, $g_{\text{DRO}^{\text{APP}}}(\mathbf{x})$, of the operational cost.

$$\textbf{DRO}^{\textbf{APP}}\text{:} \qquad \max_{\mathbf{x}\in\mathcal{X}} \quad \boldsymbol{b_0}^{\intercal}\mathbf{x} + g_{\text{DRO}^{\text{APP}}}(\mathbf{x}) \tag{5.27}$$

$$\text{where:} \qquad g_{\text{DRO}^{\text{APP}}}(\mathbf{x}) = \inf_{P\in\mathcal{D}} \mathbb{E}_P\left[f(\mathbf{x},\boldsymbol{\xi})\right] \tag{5.28}$$

The inner operational problem that computes the worst-case expected value of overstay can be expressed as in Equation (5.28). We employed historical data of the patient stay as well as expert estimates to define the support and the expected value of $\boldsymbol{\xi}$. We define a set of a compact support, Ξ, and the expected value of $\bar{\boldsymbol{\xi}}$, for $\boldsymbol{\xi} \in \Xi$. Thus, the ambiguity set, $\mathcal{D}$, in Equation (5.29), represents the family of distributions with mean value, $\bar{\boldsymbol{\xi}}$, within the support, Ξ. $\mathcal{P}$ is the set of all probability measures in the measurable space, $\mathcal{N}$ and, $\mathbb{E}_P[\boldsymbol{\xi}]$, represents the expected value of $\boldsymbol{\xi}$ under a given probability measure P.

$$\mathcal{D} = \left\{ P \in \mathcal{P} : \mathbb{E}_P[\boldsymbol{\xi}] = \bar{\boldsymbol{\xi}} \right\} \tag{5.29}$$

Considering the information about the expected value of $\boldsymbol{\xi}$, the ambiguity-averse expected operational cost in (5.28), assumes the following form:

$$g_{\text{DRO}^{\text{APP}}}(\mathbf{x}) = \inf_{P \in \mathcal{D}} \int_{\Xi} f(\mathbf{x}, \boldsymbol{\xi}) \, dP(\boldsymbol{\xi}) \tag{5.30}$$

In the next section, we solve problem (5.28) under the ambiguity-averse measure defined in Equation (5.30).

5.3. Solution methodology

In this section, we employ the scenario-based solution methodology presented in Chapter 2, to solve the DRO$^{\text{APP}}$ model.

Similar to what indicated in Subsection 2.4.1, the ambiguity-averse expectation in Equation (5.30) can be expressed as an infinite-dimensional linear optimization problem.

The typical form of the classical moment problem (Landau, 1987) is presented in Equation (5.31), where α and γ_i are the dual variables related to moment problems constraints.

$$\begin{aligned} g_{\text{DRO}^{\text{APP}}}(\mathbf{x}) = \inf_{P \in \mathcal{D}} & \int_{\Xi} f(\mathbf{x}, \boldsymbol{\xi})\, dP(\boldsymbol{\xi}) \\ \text{s.t.:} & \int_{\Xi} dP(\boldsymbol{\xi}) = 1 \qquad\qquad : \alpha \\ & \int_{\Xi} \xi_i\, dP(\boldsymbol{\xi}) = \bar{\xi}_i \quad \forall i \in I : \gamma_i \end{aligned} \tag{5.31}$$

The finite single-level equivalent formulation of problem 5.31 can be defined as follows,

$$\begin{aligned} g_{\text{DRO}^{\text{APP}}}(\mathbf{x}) = \max_{\alpha, \boldsymbol{\gamma}} \quad & \alpha + \boldsymbol{\gamma}^{\intercal} \bar{\boldsymbol{\xi}} \\ \text{s.t.:} \quad & f(\mathbf{x}, \boldsymbol{\xi}) \leq -\alpha - \boldsymbol{\gamma}^{\intercal} \boldsymbol{\xi} \qquad \forall \boldsymbol{\xi} \in \Xi. \end{aligned} \tag{5.32}$$

Note that as we explained in Subsection 2.4.2, the infinite set of constraints in problem (5.32) can be represented as $\max_{\boldsymbol{\xi} \in \Xi} \left(f(\mathbf{x}, \boldsymbol{\xi}) + \boldsymbol{\gamma}^{\intercal} \boldsymbol{\xi} \right) \leq -\alpha$. By definition, $f(\mathbf{x}, \boldsymbol{\xi})$ is a convex function on $\boldsymbol{\xi}$. Thus, expression $f(\mathbf{x}, \boldsymbol{\xi}) + \boldsymbol{\gamma}^{\intercal} \boldsymbol{\xi}$ is a convex function on $\boldsymbol{\xi}$. Therefore, the optimal value of $\max_{\boldsymbol{\xi} \in \Xi} \left(f(\mathbf{x}, \boldsymbol{\xi}) + \boldsymbol{\gamma}^{\intercal} \boldsymbol{\xi} \right)$ would be at any of the vertexes of the box type support set of $\boldsymbol{\xi}$, Ξ.

We define $\boldsymbol{\xi}_k$ as the extreme points vector of our uncertain parameter, $\boldsymbol{\xi}$ and use k to index all extreme points of the support set, Ξ. Then, the DRO$^{\text{APP}}$ model in Equations (5.27)–(5.28) can be formulated as an equivalent MILP as in (5.33).

$$
\begin{aligned}
\max_{\mathbf{x},\mathbf{y},\alpha,\boldsymbol{\gamma}} \quad & \boldsymbol{b}_0^{\intercal}\mathbf{x} + \alpha + \bar{\boldsymbol{\xi}} \\
\text{s.t.:} \quad & \mathbf{x} \in \mathcal{X} \\
& \boldsymbol{b}_1^{\intercal}\mathbf{y}_k \leq -\alpha - \boldsymbol{\gamma}^{\intercal}\boldsymbol{\xi}_k \qquad \forall k \in K \\
& \boldsymbol{W}\mathbf{y}_k \geq \boldsymbol{B}\boldsymbol{\xi}_k - \boldsymbol{G}\mathbf{x} \qquad \forall k \in K
\end{aligned}
\tag{5.33}
$$

The set of Equations (5.34) shows the equivalent robust optimization approach to solve the RO$^{\text{APP}}$ derived from the formulation in the group of Equations (5.33). The solution of the operational model will be the worst-case scenario in the support set Ξ, as we present in (5.34).

$$
\begin{aligned}
\max_{\mathbf{x},\mathbf{y},\alpha} \quad & \boldsymbol{b}_0^{\intercal}\mathbf{x} + \alpha \\
\text{s.t.:} \quad & \mathbf{x} \in \mathcal{X} \\
& \boldsymbol{b}_1^{\intercal}\mathbf{y}_k \leq -\alpha \qquad \forall k \in K \\
& \boldsymbol{W}\mathbf{y}_k \geq \boldsymbol{B}\boldsymbol{\xi}_k - \boldsymbol{G}\mathbf{x} \qquad \forall k \in K
\end{aligned}
\tag{5.34}
$$

5.4. Numerical studies

The case study used to validate our proposed methodology is based on a public health center in Chile. The hospital is a national center of reference and receives about 11,000 requests of admission yearly. The institution has ten care units (i.e., areas of allocation) and serves nearly twenty different diagnosis groups. We focus on the process of admission to ward beds, which is managed from the CAD that makes daily decisions of scheduling and allocation. The evaluated case study describes the prevalent situation in the hospital under study. The monthly admission plan in the CAD is executed manually, and there is no

formal method or standard policy of admission as it happens in most hospitals (Gemmel & Van Dierdonck, 1999).

This section describes the numerical studies of the proposed approach. Subsection 5.4.1 describes the performance metrics to be evaluated and Subsection 5.4.2 explains the data requirements.

5.4.1. Performance metrics

In order to assess the performance of the proposed approach, we defined several metrics. These metrics allow for measuring the level of service (AP_s) and hospital performance (RU_{rt}, MO_{is}). Detailed definition of the metrics is given as follows:

(i) *Percentage of admitted patients*: This metric measures the total percentage of admitted patients by type, s.

$$AP_s = \frac{\sum_{r=1}^{R}\sum_{i=1}^{I} x^{\mathrm{P}}_{ris}}{\sum_{i=1}^{I} D_{is}} \times 100 \quad \forall s \in S \tag{5.35}$$

(ii) *Percentage of resource utilization*: This metric measures the total percentage of beds occupied per room, r, and period, t.

$$RU_{rt} = \frac{\sum_{i=1}^{I}\sum_{s=1}^{S} x^{\mathrm{S}}_{rist}}{C_{tr}} \times 100 \quad \forall r \in R, t \in T \tag{5.36}$$

(iii) *Percentage of maximum budget of overstay utilization*: This metric measures the total percentage of overstay days used, per patient, i, and type, s, over the total allowed.

$$MO_{is} = \frac{\sum_{r=1}^{R} o_{ris}}{\sum_{r=1}^{R} O_r^{max}} \times 100 \quad \forall i \in I, s \in S \tag{5.37}$$

5.4.2. Data description

The numerical experiments are based on real data from the EHR of the hospital under study[1]. The database accounts for information on the inpatient service for the period 2010-2016. It includes data about the patient's date of admission and discharge, admission care unit, length of stay, diagnosis, and DRG and SII indexes. Below, we describe the characteristics and setup of the parameters used in the study. The input data of the model parameters is detailed in Appendix B.

A. Demand

The CAD receives requests for admission from different care units in the hospital. We characterize the demand, D_{is}, as the number of requests of elective patients that differ by type, i.e., diagnosis. We remark that the admission requests are in terms of the "number of patients." Hence, considering that our approach accounts simultaneously for scheduling and allocation decisions, the model schedules patients and allocates days of stay. For the case study, we considered demand requests of sixty patients, $I = 60$, that belong to twelve different patient diagnoses[2], $S = 12$, to be allocated in ten room types, $R = 10$, and a time horizon of $T = 30$ days[3]. Table 5.2 summarizes the patient type classification according to the DRG. A description of the diagnosis can be found in Laguna et al. (2000). The table also shows the historical rates of admission for the care units with greater patient flow, i.e.,

[1]Specific details about the hospital under study have been omitted for the sake of privacy.
[2]Note that to each patient, i, is associated a type, s, according to their diagnosis.
[3]In Appendix B, Table B.3 we detail the input data of the patient admission.

86% of the total admissions in the CAD (Medical 37%, Surgical 41%, Ophthalmology 8%). Based on such proportions, we determined the distributions of patients scheduled in the corresponding care units. For example, in the Ophthalmology unit, only the patient types C, H, S, and Z, can be assigned. An extended version of Table 5.2 is presented in Appendix B, Table B.2.

TABLE 5.2. Mean value (in days), support set (in days) and proportion of admission of patient type per care unit from period 2010-2016.

	Length of Stay Data		% of admission request per care unit		
Diagnosis	ξ	Ξ	Medicine	Surgery	Ophthalmology
C	9.93	[2,26]	7.6%	12.9%	0.7%
F	13.02	[2,27]	1.1%	0.0%	0.0%
H	5.06	[1,15]	0.6%	0.4%	43.0%
I	11.26	[1,26]	29.0%	6.3%	0.0%
K	6.71	[1,21]	8.1%	22.1%	0.0%
S	9.18	[1,25]	3.0%	21.9%	48.5%
Z	5.26	[0,12]	3.4%	1.1%	1.7%

B. Capacity availability

The hospital under study is divided into several care units according to the clinical specialty, which are subdivided into room types. We define as "*room*" the aggregate set of rooms associated with a care unit. The beds that belong to each room are identical. The capacity availability, C_{tr}, is defined as the total number of available beds per time period and room. We employed historical data to determine bed availability in the time horizon. In our study, we computed the expected value of daily bed availability per room for the period 2010-2016. Such information can also be estimated by employing machine learning techniques, as stated in (Bertsimas, Pauphilet, et al., 2019). We remark that this data corresponds to the total daily available capacity, including the beds for unscheduled patients.

Figure 5.2 illustrates an example (from real data) of the pattern of daily bed availability for a typical month for the three main care units of allocation. We note that due to the time-varying bed availability, ensuring continuity in allocation is difficult to achieve. This high variability could be attributed to hospital planning deficiencies (e.g., discharge), which should be revised to guarantee better admission levels.

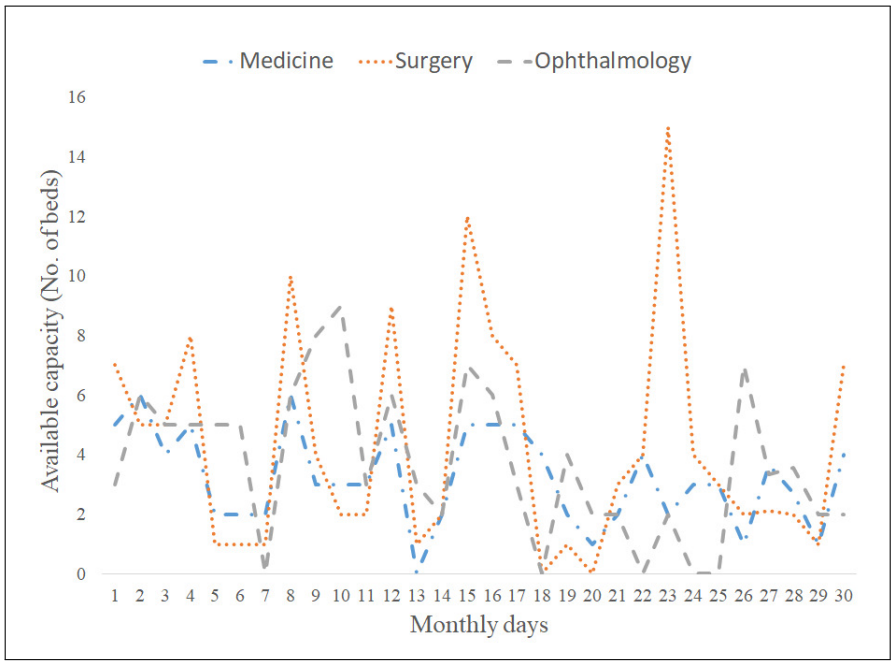


FIGURE 5.2. Daily bed availability graph by care unit in the hospital (typical month).

C. Patient Length of Stay

We consider the patient LoS, $\boldsymbol{\xi}$, as the uncertain parameter. Figure 5.3 details the statistics (from real data) of the patient's length of stay by type of diagnosis. We observe not only a high variability in the data but also a difference in the distribution between patient types. The input data for the case study is reported in Table 5.2. It shows the mean, $\bar{\boldsymbol{\xi}}$, and support, Ξ (main diagnoses), per patient type, obtained from historical data. The support set was defined considering data between the 5th and 95th percentile, i.e., we account for 90% percent of the data distribution.

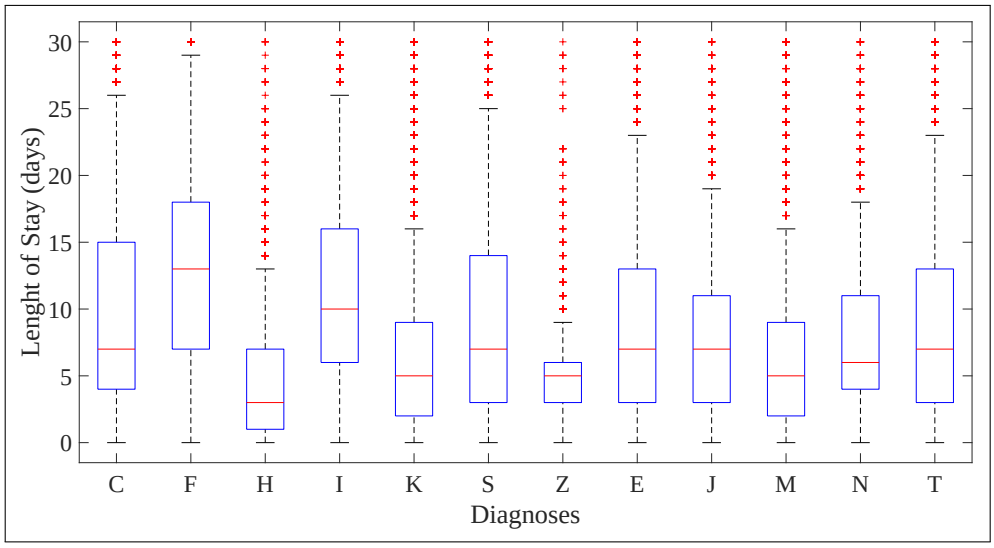


FIGURE 5.3. Length of stay distributions of patient type (diagnosis) from period 2010-2016.

D. Benefit of admission

In practice, health care providers decide the admission of patients in terms of their health priority. From the EHR, we employed the DRG and the SII indexes to estimate the weighted priority of the patients, ζ_{is}. We assumed that a high DRG score indicates that the patient has a higher priority. The SII ranks the patients between three levels of illness acuity; Major (3), Moderate (2) and, Minor (1). We developed a statistical analysis to calculate the correlation between patient severity illness and DRG to determine the association between the variables, employing a parametric test of Pearson correlation. The results of the analysis indicated a significant correlation at a significance level of $P < 0.05$ (2-tailed), between the DRG and SII indexes (See Table B.1 in Appendix B). Considering the reported results, we can assume that the DRG weight is an adequate score to characterize the patient's priority.

E. Cost of overstay

The cost of overstay, θ_{is}, refers to the penalty for shortening the patients stay duration in the allocation process. The cost can be seen as a detriment of patient service, as the patients will have to be transferred or wait to be allocated, among other actions. In practice, the weights have to be defined by hospital managers based on particular preferences and internal policies. To set up the weights of overstay, we assumed proportional values of the benefit of allocation, which varies according to the patient type.

F. Maximum budget of overstay

In order to define the maximum number of overstay days per room, r, we have included the parameter, O_r^{max}. The value should be set by managerial decisions and will depend on the level of service that is expected to be offered. For this study, we decided to choose a value arbitrarily and test its influence over the optimal admission decisions through a sensitivity analysis. We defined a total threshold of maximum overstay of, $O_r^{max} = 300$ days, which differs according to the room.

5.4.3. Results and discussion

In this section, we present the computational experiments along with the results and discussions of the case of study. We first compare the in-sample solutions in terms of the scheduling and allocation decisions of the distributionally robust model (DRO^{APP}) with the solutions of standard models, robust model (RO^{APP}), stochastic model (SP^{APP}), and, deterministic model (DT^{APP}), in Subsection 5.4.4. In Subsection 5.4.5, we perform an out-of-sample analysis of the benchmark models to compare the results in terms of cost and reliability. In Subsection 5.4.6, we present a sensitivity analysis over the maximum overstay budget. Finally, in Subsection 5.4.7, we analyze the computational performance of the DRO^{APP} model and its capability to solve large instances.

The models have been implemented in CPLEX 12.7.0, with an optimality gap set as 0.0001%. All computations were performed on a computer Intel Xeon Gold 6148, 2.4GHz - 2.39GHz (2 processors) with 256 GB of RAM.

5.4.4. Benchmark analysis

The proposed DRO$^{\text{APP}}$ model is compared with the benchmarks approaches RO$^{\text{APP}}$, SP$^{\text{APP}}$, and DT$^{\text{APP}}$. The RO$^{\text{APP}}$ model is solved as we defined in Equations (5.34). It accounts for the support, Ξ, of the uncertain parameter, $\boldsymbol{\xi}$, ignoring the moment information. The DT$^{\text{APP}}$ model is solved by considering the nominal values of the uncertain parameter, obtained from the support set, Ξ, and the moment information, $\bar{\boldsymbol{\xi}}$. The SP$^{\text{APP}}$ model, (5.23)–(5.24), is solved by using a SAA method (Birge & Louveaux, 2011). By doing so, we generated 1000 iid random samples, assuming a known distribution fitted using historical data. We remark that we assumed different probability distributions depending on the patient type, ξ_{is}.

The in-sample results of the four optimization models are listed in Table 5.3. Column 2 shows the average percentage of admitted patients over all the patient types, (AP_s). Column 3 indicates the average percentage of bed utilization for each time period, (RU_{rt}). Column 4 indicates the average percentage of the maximum budget of overstay utilization, (MO_{is}). Column 5 reports the first stage results related to the benefit of admission and, column 6 shows the total benefit, including the cost of overstay. Columns 7–10 reports the computational times and size of the models. Overall, the computational time of the DRO$^{\text{APP}}$ is shorter than the SP$^{\text{APP}}$ and RO$^{\text{APP}}$ models; it takes 45 minutes to solve to optimality, which is acceptable for weekly to daily scheduling. In contrast, the SP$^{\text{APP}}$, which is an approximate model with finite samples, takes about 75 minutes. The computational time in DT$^{\text{APP}}$ model is shorter than the other models, as a result of the difference in size.

TABLE 5.3. Admission Planning Problem In-sample solutions.

	In-sample solutions					Computational time / Size			
Model	Admitted patients (AP_s %)	Bed utilization (RU_{rt} %)	Maximum budget (MO_{is} %)	Admission benefit	Total benefit	CPU time (s)	Constraints	Binary variables	Linear variables
RO^{APP}	25.15	23.78	4.17	8.87	7.51	1861.77	252231	14539	247466
DRO^{APP}	44.40	37.65	32.33	29.34	19.32	1780.21	252231	14539	247526
SP^{APP}	57.50	41.31	18.47	41.79	34.89	2699.37	146923	15869	61705
DT^{APP}	85.69	34.09	38.85	56.41	47.80	1.30	14723	14539	1765

From Table 5.3, it can be seen that the solutions of the RO^{APP} model are the more conservative in terms of the percentage of admitted patients, (25.15%), average bed utilization, (23.78%), and percentage of overstay budget use, (4.17%), which implies a lower benefit of admission in comparison to the other approaches. The DRO^{APP} solutions are less conservative than the RO^{APP}, mainly because it incorporates partial information omitted by the RO^{APP}. Compared to the DRO^{APP}, the SP^{APP} model gives a higher value of admission rate, for an increment of 18.33% of the percentage of admitted patients and 4% in bed utilization. The results can be explained because the DRO^{APP} model seeks to hedge against the worst-case probability distribution in the ambiguity set. Therefore, it protects itself against the inaccuracy of probability distributions under all ambiguities. Interestingly, we observe that the percentage of overstay use for the SP^{APP} model is lower than the DRO^{APP} and DT^{APP} model's values. Such results suggest that the SP^{APP} model, which assumes expected values of the patient's LoS, focuses on scheduling a high number of patients with mean values of time allowances, that will require shorter overstay budget use. The DT^{APP} model has similar behavior to SP^{APP} admitting the highest rate of patients, 85.69%, but with a high percentage of the overstay budget, 39%, to compensate for the short stay durations allocated.

We remark that according to historical data of the hospital under study, the average percentage of monthly scheduled elective patients (AP_s) is approximately 11%, implying a lack of visibility of future events, which lead to a pessimistic decision-making approach. We can state that the results of the proposed DRO approach outperform the hospital's

actual practice. In the next subsection, we extend the analysis of in-sample solutions regarding the scheduling and allocation decisions.

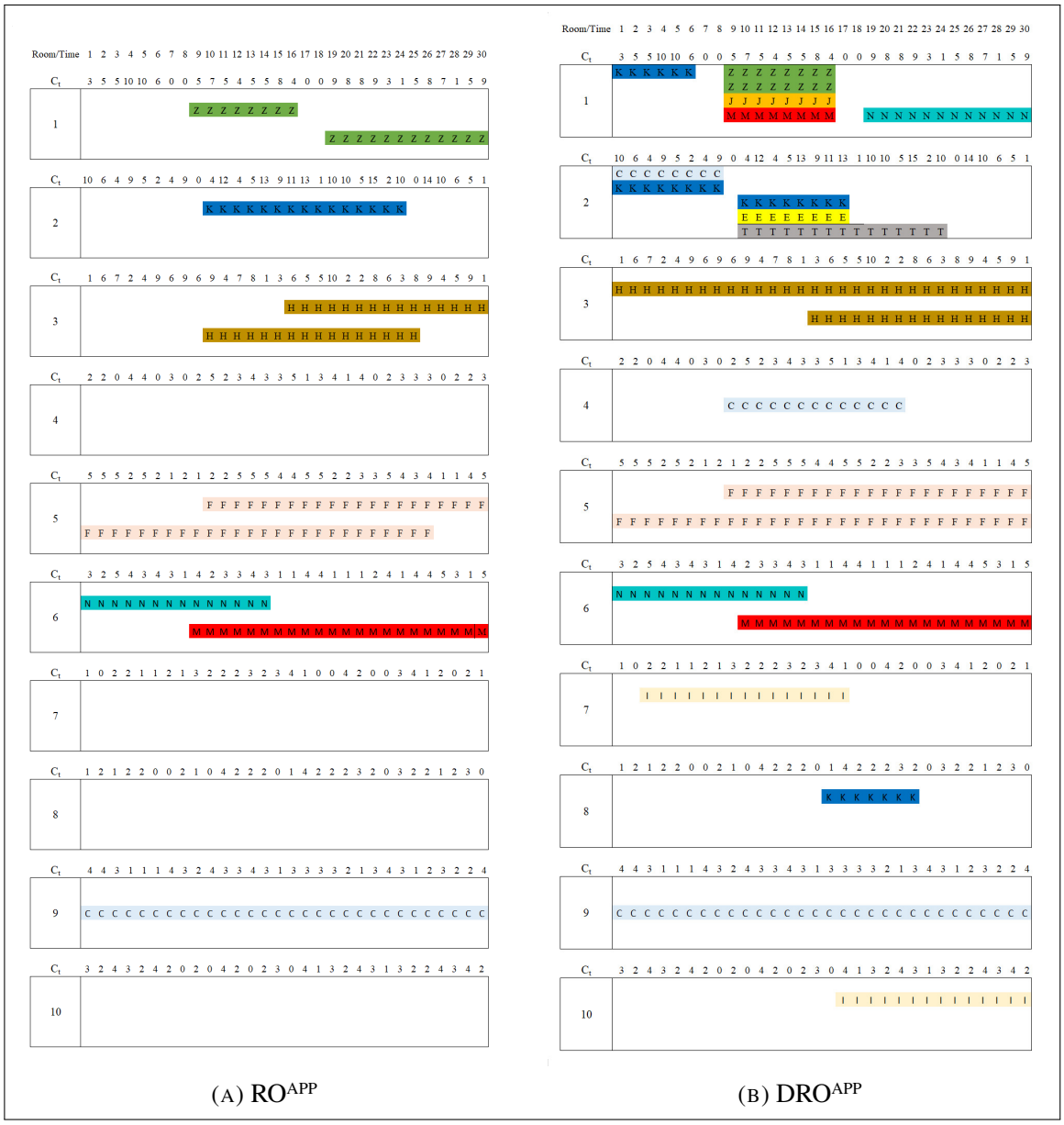
5.4.4.1. Scheduling and allocation decisions

In this subsection, we analyze and compare the tactical admission solutions of the evaluated models, in terms of the reserved days of stay, $\mathbf{x}^{\mathrm{S}}$, the overstay, $\mathbf{o}$, and, the effect of the time-varying capacity availability, C_{tr}, over the scheduling-allocation decisions.

The optimal schedule for the four evaluated models is compared in Figure 5.4. The scheduling plan indicates the start times, $\mathbf{x}^{\mathrm{A}}$, reserved days, $\mathbf{x}^{\mathrm{S}}$, and assigned room of patient type, s. Each line in the timetable represents a patient, i, on the waiting list. The figure also includes the vector of available capacity, C_{tr}, per time period, and room. We remark that due to we account for a patient-to-room approach, the sequencing problem for individual patients is not evaluated. We instead study the reserved days of stay[4], per patient type.

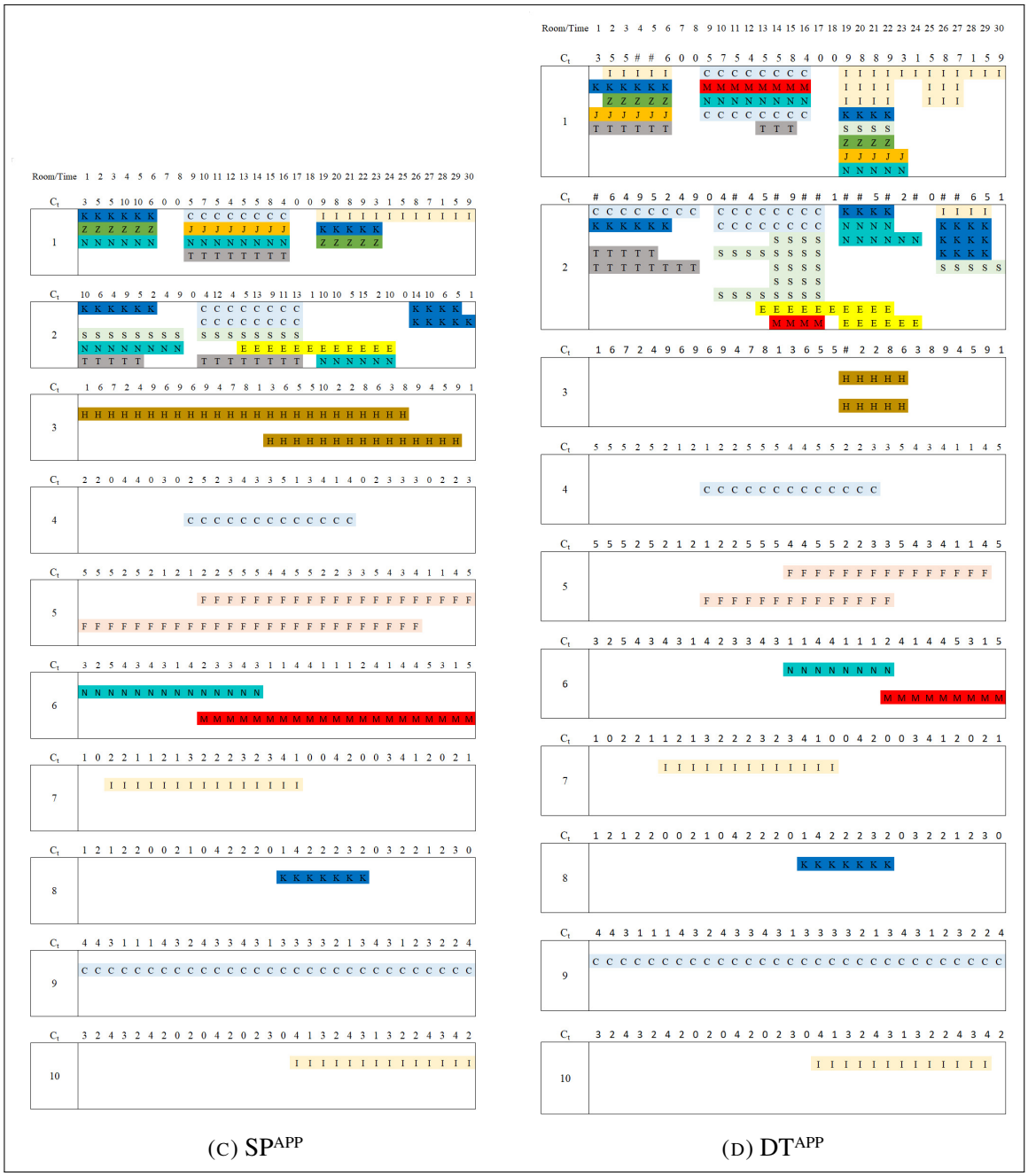
The solutions of the benchmark models can be analyzed from several angles, such as the decisions about the reserved days of stay, the day of admission by patient type, the number of admitted patients. Those aspects are determined by the level of conservatism of the models, whether they are ambiguity-averse or risk-neutral. The RO$^{\text{APP}}$ model, Figure 5.4a, admits fewer patients per room with longer days of stay compared to the other approaches. Besides, the admission date of the patients varies for some of the rooms. In contrast, the DRO$^{\text{APP}}$ model, Figure 5.4b, keeps a balance in the number of admissions and the reserved days of the patient's type; it admits more patients with shorter time allowances than the RO$^{\text{APP}}$ model.

[4]Some studies refer to as time allowances.



(A) RO^{APP} (B) DRO^{APP}

FIGURE 5.4. Optimal schedule of patients of the RO^{APP} (A), DRO^{APP} (B), SP^{APP} (C) and DT^{APP} (D) models.



(C) SP^{APP} (D) DT^{APP}

FIGURE 5.4. Optimal schedule of patients of the RO^{APP} (A), DRO^{APP} (B), SP^{APP} (C) and DT^{APP} (D) models (cont.).

Interestingly, the SP^{APP}, Figure 5.4c, and DT^{APP}, Figure 5.4d, models, admits a more significant number of patients with shorter reserved days. For these cases, the admission delays are smaller, i.e., more patients are admitted at the beginning of the horizon, compared to the ambiguity-averse approaches. However, as we explained in Subsection 5.4.5, the solutions will result in high regret due to the risk-neutral approach in which the expected values of the LoS are considered.

Three main insights can be obtained from Figure 5.4; firstly, the admission date of certain patients could be delayed in the time horizon; secondly, there exist under-utilization between periods, in which the capacity exceeds the bed requirements. The timetable underlines the effect of time-varying bed capacity and LoS uncertainty in the scheduling plan. Finally, the timetable also shows that the optimal time allowances per patient type can differ by lengths and patterns, depending on the capacity availability in the time horizon. Therefore, the intuitive safety factor pattern of allocation *"mean plus safety stock"* presented in Mak et al. (2014) is not accomplished in this particular case under time-varying capacity. Below we explain those aspects in detail.

Reserved days of stay. The reserved days of stay (i.e., time allowances), x^S, for the four evaluated models are compared in a box-plot graph, Figure 5.5. The x-axis indicates the patient diagnoses, and the y-axis denotes the reserved days of stay. We observe that the RO^{APP} model, Figure 5.5a, schedule fewer patients types with conservative values of time allowances. The DRO^{APP} model, Figure 5.5b, allocates longer time allowances resulting in better use of resources. The solutions are less conservative, on average, and schedules a more significant number of patients types. The stochastic model, SP^{APP}, Figure 5.5c, which accounts for perfect information of the LoS distribution, assign all the patient types while assuming expected values of the stays. Finally, contrasted to the previous cases, the

deterministic model, DT$^{\text{APP}}$, Figure 5.5d, allocates significantly more patients with shorter time allowances.

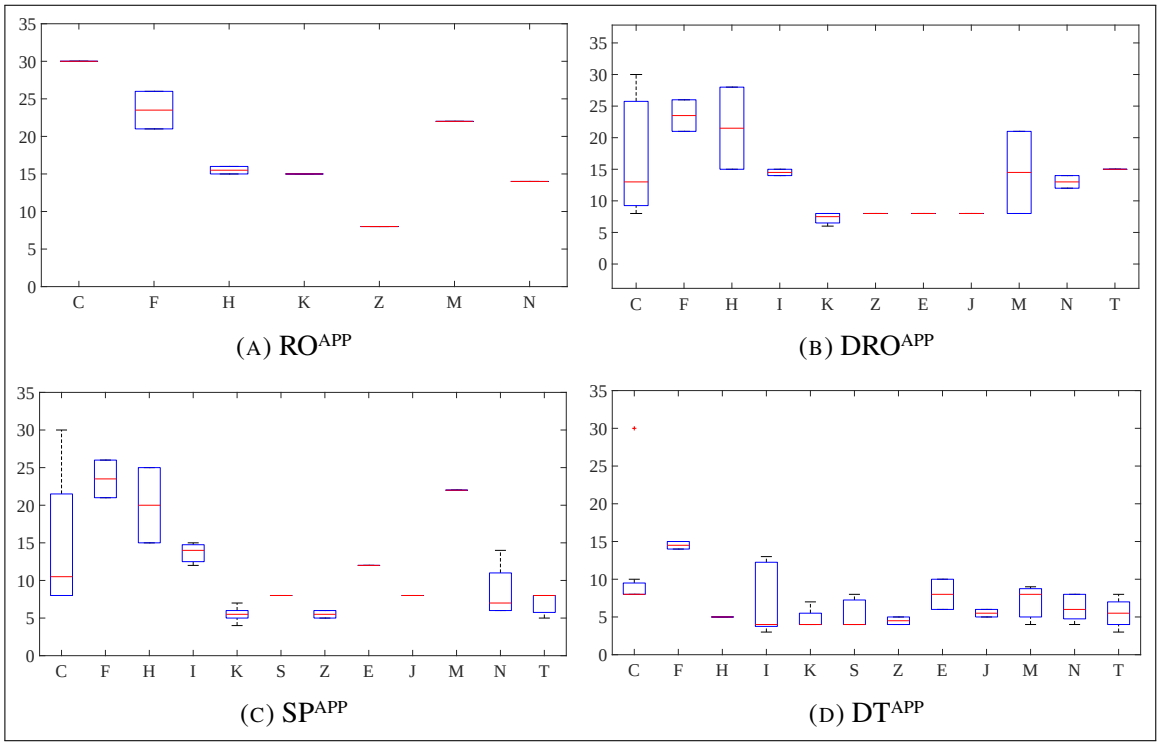


FIGURE 5.5. Box-plot representation of the time allowances per patient type for the evaluated models.

From the results, we can infer that a decision-maker who is ambiguity-averse will prefer to schedule fewer patients with longer reservations of stay, to avoid overstay costs due to the LoS variations. For example, for the patient type K, the DRO$^{\text{APP}}$ model reserves on average, 7.25 days; the SP$^{\text{APP}}$, 5.33 days; the DT$^{\text{APP}}$, 4.43 days. Recall that from the historical data (see Table 5.2), the mean LoS for this patient type is $\bar{\xi}_K = 6.71$ and the support, $\Xi = [1, 21]$. Accordingly, the average length of reserved days in the SP$^{\text{APP}}$ and DT$^{\text{APP}}$ models are less than the mean LoS of patient K, which could result in infeasible solutions at the operational level.

***Overstay - Early discharge policy*.** The early discharge policy aims to offer a certain degree of flexibility in the admission process due to the uncertain LoS. In that sense, since we account for a multi-specialty and multi-priority framework, it is important to evaluate the resulted admission case mix of the proposed approach, i.e., which patients are admitted according to their priority and early discharge days. This analysis extends the results presented in Table 5.3, about the percentage of the maximum budget of overstay days used, MO_{is}, for the evaluated models.

Figure 5.6 illustrates the admission plan for the evaluated models, as a function of the patient priority (i.e., DRG), ζ_{is}, on the horizontal axis, and the early discharge days assigned per patient, o_{ris}, on the vertical axis. We note that there is not a clear association between the patient priority and allocated days of overstay. This result indicates that early discharging only the lower priority patients may not be the best admission policy. The DT$^{\text{APP}}$ model admits a more significant number of patients in the priority range $[1 - 1.5]$, compared with the other models in which the highest admission frequency is in the priority range $[0-1]$. We also note that the allocated early discharge days vary among the evaluated models according to the patient priority weight. The DT$^{\text{APP}}$ and DRO$^{\text{APP}}$ models consider early discharge days in the scale range $[0-9]$. On the contrary, the SP$^{\text{APP}}$ and RO$^{\text{APP}}$ models allocate early discharge days in a lower scale range $[0 - 4]$.

By comparing these solutions with the decisions about the reserved days of stay, and the proportion of admitted patients of the evaluated models, we observe that the results are highly dependent on the level of protection to future changes and the assumptions made over the uncertain LoS. The DRO$^{\text{APP}}$ model reserves longer stay durations due to its ambiguity approach over the distribution of the patient length of stay, which causes the allocation of higher values of early discharge days. The DT$^{\text{APP}}$ model admits a higher number of patients with very short stay reservations due to its myopic approach, which

requires a high use of early discharge days. In contrast, the SP^{APP} model consumes fewer overstay days as a result of assuming expected values of the LoS. Finally, the RO^{APP} model is the more conservative, and longer stay durations are reserved that require fewer days of overstay to protect itself to future changes. Overall, this analysis reveals that tactical admission decisions are not trivial; scheduling only patients with the highest priority will not guarantee better performance in terms of the number of admitted patients and bed utilization.

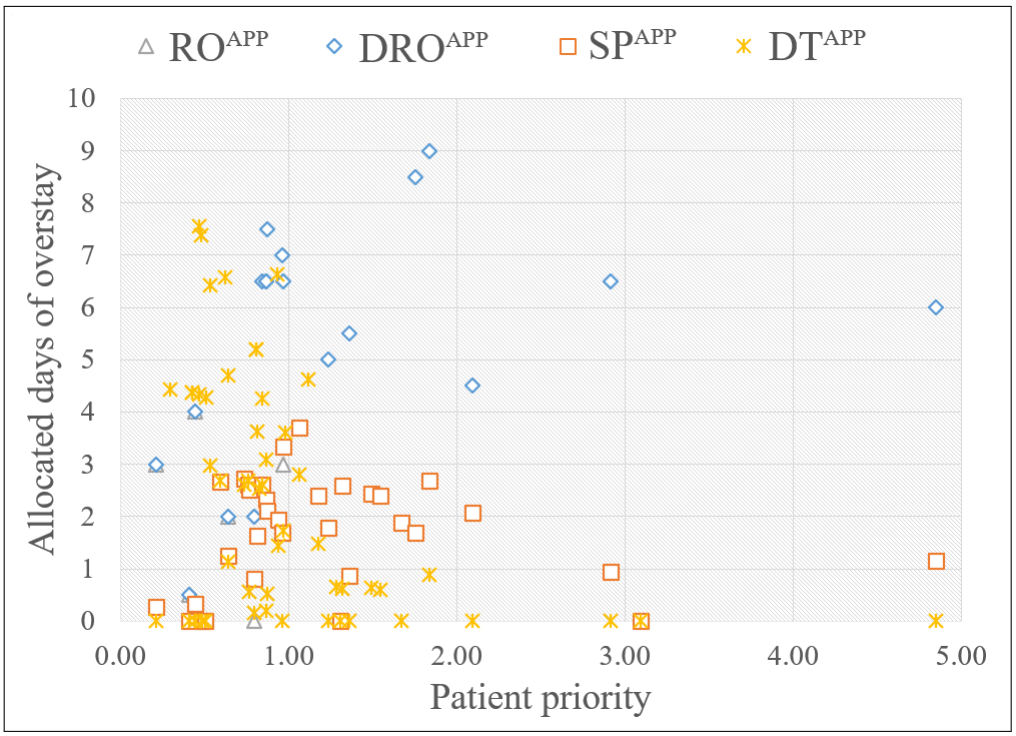


FIGURE 5.6. Comparative of the admission case mix as a function of the patient priority weight and the allocated overstay days for the evaluated models.

Capacity availability. Figure 5.7 compares the solutions of the average daily bed utilization among the evaluated models. The SP^{APP} and DRO^{APP} models show the highest levels of bed utilization, while the RO^{APP} model indicates the lowest levels due to its over-conservative approach. The DT^{APP} model, which reserves short stay durations, has low

utilization levels, allocating fewer patients at the beginning of the time horizon (i.e., t = 1 – 13) compared to the other models. It can also be observed that the bed utilization decrease at the end of the time horizon for all cases, due to the increase in available capacity in such periods.

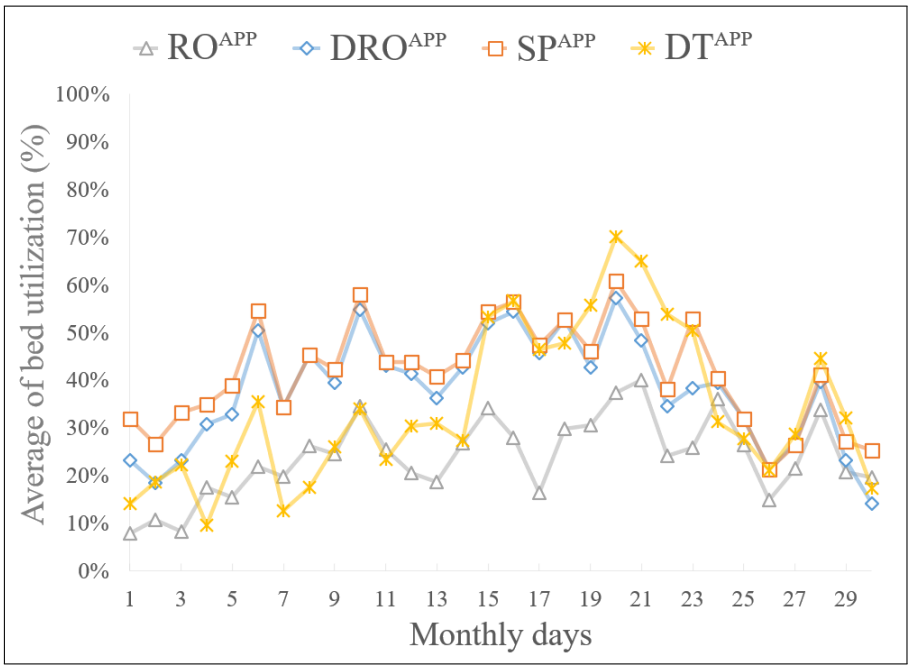


FIGURE 5.7. Comparative results of the daily bed utilization of the APP of the evaluated models.

Interestingly, the resulting daily occupancy rates are in the range of 10% – 60% compared to the target of 85% for a medium-large size hospital. Such results can be explained by remarking that the historical data used about bed availability includes the total daily availability in the hospital, including the beds for unscheduled patients; this slack capacity corresponds to 20% of the total availability. Besides, the standard measure of 85% of occupation does not capture the variations in bed occupancy throughout the day, which causes the general thought that hospitals have excess capacity (Green, 2002). Thus, according to the reported results, we can assert that this perception is the consequence of

the inherent daily variations in capacity availability and uncertain length of stay, making it more difficult to guarantee the highest levels of bed occupancy. Nevertheless, the proposed tactical-operational plan may allow decision-makers to anticipate the planning and adjust it with an appropriate supply of beds, in which the idle capacity could be used as a buffer to allocate unscheduled patients.

5.4.5. Out-of-sample analysis

In order to make fair comparisons and to measure the performance of the solutions of the evaluated models (RO^{APP}, DRO^{APP}, SP^{APP}, DT^{APP}), we performed an out-of-sample test.

Table 5.4 reports the out-of-sample solutions. We performed four out-of-sample experiments, named as MIX PROB 1–4. In contrast to most studies in the literature, that assume exponential or log-normal distributions for the patient LoS (He et al., 2019), in our study, we tested the solutions considering different probability distributions according to the patient's diagnoses, fitted using historical data. The first three experiments consider a mix of the LoS probability distribution differing by patient type, and as a mode of comparison, for the fourth case, the LoS is assumed exponential for all patient types. It should be noted that, since we are working with a real dataset, the fitting data of the patient's LoS accounts for high-variance distributions with a coefficient of variation over 80% for all of the cases. Hence, we are evaluating the models for extreme instances of uncertainty. In Appendix B, Table B.5, we detail the parameters and probability distributions of each experiment.

We generated 10000 independent and identically distributed random samples by patient type to perform the experiments. The operational cost of patient overstay was computed for each scenario by employing the framework scheme based on Equations (5.33) and (5.34) for the DRO^{APP} and RO^{APP} models, respectively. For the SP^{APP} model, we

considered the SAA method as explained in Subsection 2.2.2, Equation (2.6). In order to assure feasibility, the test was implemented by relaxing the hard constraint (5.18).

To evaluate the reliability and robustness of the solutions, we defined a reliability index, RI, computed over the 10000 scenarios of the test data. The RI indicates a frequency of occurrence; it assesses the feasibility of the model by calculating the average frequency in which the overstay days in the out-sample solution are less than or equal to the solutions obtained in the in-sample analysis.

TABLE 5.4. Admission Planning Problem out-of-Sample solutions.

Model	MIX PROB 1		MIX PROB 2		MIX PROB 3		Exponential distribution	
	Total Benefit	RI (%)	Total Benefit	RI (%)	Total Benefit	RI (%)	Total Benefit	RI (%)
RO^{APP}	8.60	94.80	8.53	94.43	8.37	94.38	8.30	92.05
DRO^{APP}	24.28	91.77	24.89	92.58	22.70	90.38	22.63	87.24
SP^{APP}	31.58	73.64	32.51	74.44	29.15	72.64	29.33	69.64
DT^{APP}	35.56	47.64	37.43	49.46	33.36	49.72	33.75	49.66

From Table 5.4 several observations can be derived. We note that the solutions are similar for the first three experiments despite considering a mix of patient types distributions. Nevertheless, the fourth experiment's solutions in which we assumed an exponential distribution for all patient types have the lower performance for the RO^{APP}, DRO^{APP}, and SP^{APP} models, compared to the other experiments. Therefore, we can state that considering various probability distributions of the patient's LoS, differing by patient type, assures the admission plan's robustness at the tactical level. The DRO^{APP} approach exhibits good performance for both the total benefit of admission and the RI among all tested experiments. When compared to the RO^{APP} model, the solutions are less conservative with similar values of the RI. Note that as indicated in Table 5.3, the DRO^{APP} model guarantees 57% more admission rate than the RO^{APP} model, for only a 3% difference

in the RI. In contrast, the SP$^{\text{APP}}$ and DT$^{\text{APP}}$ models show a lower performance of the RI, differing in more than 40%, compared to the DRO$^{\text{APP}}$ model.

In summary, a decision-maker with limited information about the LoS distributions will obtain better performance if considering the DRO$^{\text{APP}}$ approach. The regret of using the SP$^{\text{APP}}$ and DT$^{\text{APP}}$ models is high, given that the solutions show low-reliability index and benefit of admission, in the face of extreme realizations of the patient's LoS. Hence, there will be high operational costs, which will cause transfers, delays, and appointments overlap. A more conservative approach could also be considered, as the RO$^{\text{APP}}$ model; however, this may result in low admission quotas and higher under-utilization rates of the bed resources.

5.4.6. Sensitivity analysis of the overstay maximum budget

We performed two what-if analyses considering the penalized days of overstay, O_r^{max} (i.e., overstay maximum budget). The parameter which must be specified by a decision-maker is interpreted as the total early discharge days allowed per room. For the evaluated models, we first analyze in Subsection 5.4.6.1, the number of admitted patients metric for different values of the budget of overstay. Then in Subsection 5.4.6.2, for the DRO$^{\text{APP}}$ model, we evaluate the trade-off between the maximum budget of overstay over the number of admitted patients, resource utilization, and the reliability index, RI.

5.4.6.1. Comparative sensitivity analysis of the evaluated models

Figure 5.8 shows the solutions to the analysis. For the evaluated models, we compare the average number of admitted patients (x-axis) for different values of the maximum budget of overstay (y-axis). As expected, we observe that as the maximum value of the budget of overstay increases, more patients are admitted. However, the number of admitted patients differs according to the decision approach adopted.

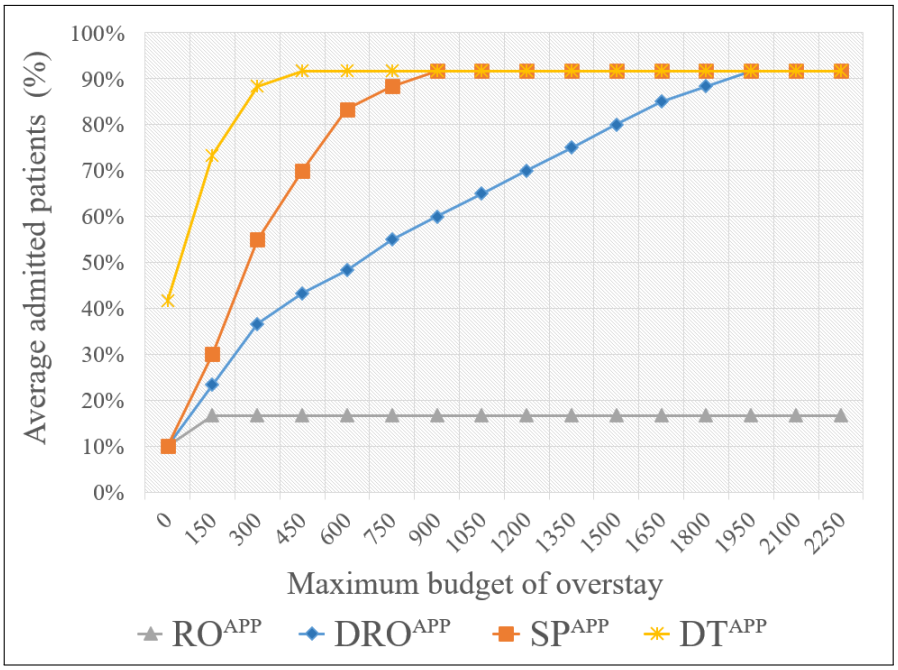


FIGURE 5.8. Comparative of the maximum budget of overstay versus the performance metric (%): admitted patients, AP_s.

The DT$^{\text{APP}}$ model, which disregards the risk in future random realizations of the uncertain LoS, considers the higher number of admitted patients (40%) for the case in which early discharge days are not allowed ($O_r^{max} = 0$). This value increases to reach a steady state due to bed capacity limitations. The RO$^{\text{APP}}$ model keeps the same admission percentage independently of the flexibility in the number of overstay days allowed; the model protects itself to a fixed value to avoid infeasibilities at the operational level. A higher level of protection is observed in the SP$^{\text{APP}}$ model, which similar to the DRO$^{\text{APP}}$, and RO$^{\text{APP}}$ models consider a percentage of admission of 10% when early discharge days are not allowed. For the SP$^{\text{APP}}$ model, the number of admissions increases rapidly as more flexibility is allowed in the admission process.

A different pattern is observed in the DRO$^{\text{APP}}$ model, which clearly protects itself from future infeasibilities but not over conservative as the RO$^{\text{APP}}$ model. This insight complements the results of the out-of-sample evaluation in which the DRO$^{\text{APP}}$ model

achieved the most effective combination of robustness and consistency, measured through the RI metric. In the next subsection, we analyze the performance of the DRO^{APP} model, including the resource utilization metric and the RI metric for different values of the maximum budget of overstay.

5.4.6.2. Sensitivity analysis evaluation of the DRO^{APP} model

Figure 5.9 exhibits the solutions of the sensitivity analysis for the DRO^{APP} model. The x-axis corresponds to the maximum budget of overstay, O_r^{max}, and the y-axis denotes the average percentage of admitted patients (left) and the average rate of bed utilization (right). The results, as shown in Figure 5.9, indicate the positive association between the metrics; as the value of maximum overstay days increases, the admission rate also increases as well as the bed utilization, but at the cost of overstay.

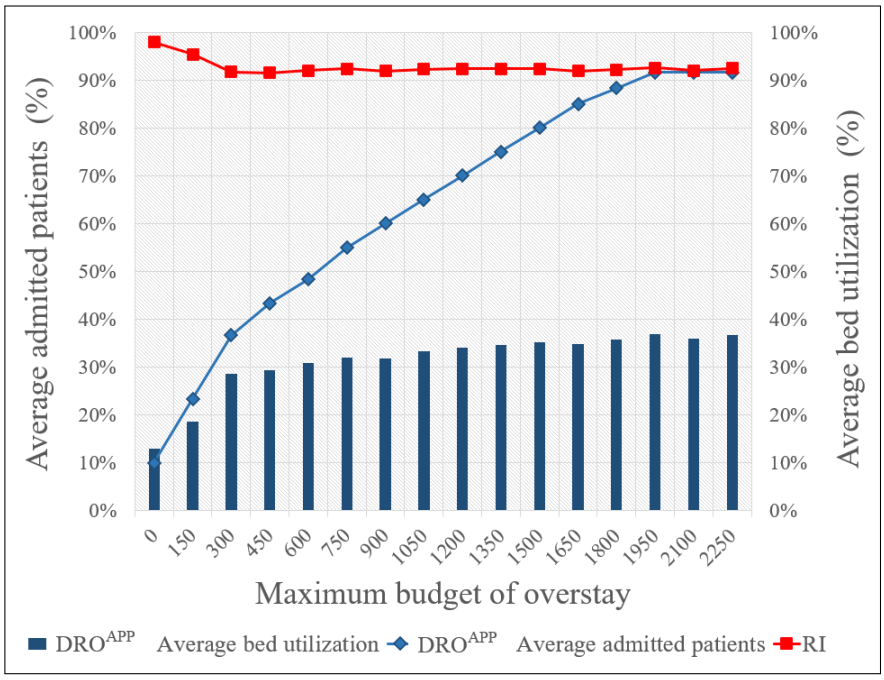


FIGURE 5.9. DRO^{APP} model comparative of the maximum budget of overstay versus the performance metrics (%): average admitted patients, AP_s, average resource utilization, RU_{rt}, and reliability index, RI.

As mentioned in the previous subsection, for the case in which early discharge days are not permitted, the model indicates a maximum percentage of admission of 10%. Hence, with an increment in the budget of overstay of 6%, it can be obtained, on average, 27% more admissions and a 16% improvement in bed utilization. It can also be observed that the solutions of the RI, reaches a steady-state as from 300 days of discharge allowed. This result can be explained due to the time-varying bed availability, which imposes an upper bound over the used early discharge days and, therefore, the number of admitted patients.

5.4.7. Computational performance evaluation

The admission planning problem has a complex combinatorial structure. The number of constraints and variables could increase exponentially according to the cardinality of patients types and rooms. We performed an analysis to test the computational performance of the proposed DRO$^{\text{APP}}$ model and its capability to solve large instances. We remark that the DRO$^{\text{APP}}$ model is solved by considering an exact MILP deterministic reformulation, based on full vertex enumeration rather than a sampling approach. Thus, as we explained in Section 5.3, we employed only the k-th extreme points in the support, Ξ, to map the scenarios, which reduces computational time. Table 5.5 compares the performance in terms of CPU time and resolution gap (%) of six configurations. We set a time limit of 24 hours to obtain the solutions.

As shown in Table 5.5 seven cases are compared. The cases correspond to different sizes of waiting lists (demand) below and above the base case (instance No. 2). Hence, the higher the number of patients, the number of rooms, and diagnoses increases. For all of the cases, we have considered a time horizon of $T = 30$ days. We observe that the DRO$^{\text{APP}}$ model can produce feasible solutions for instances of significant size in a reasonable computational time (i.e., less than 24 hours), with gap values within 1%, which

is acceptable in hospitals for tactical-operational planning. However, improvement is still needed to solve larger instances (i.e., above 14 diagnoses and 200 patients) by investigating more efficient solution methods.

TABLE 5.5. Comparative of performance evaluation DRO^{APP} model.

Instance characteristics				Computational time / size				
Instance	Patients, i (Total)	Rooms, r	Diagnoses, s	Scenarios, ξ_k	Variables (thousands)	Constraints (thousands)	CPU time (s)	Gap (%)
1	60	6	7	2187	133.30	143.31	2488.20	0.00%
2	60	10	12	4096	262.22	252.36	1780.21	0.00%
3	100	6	7	2187	238.01	247.43	7903.04	0.00%
4	100	6	12	4096	426.10	433.39	81476.38	0.64%
5	100	10	12	4096	441.81	449.06	71710.68	0.00%
6	200	10	12	4096	870.76	870.90	86062.11	1.31%
7	200	10	14	16384	3337.05	3355.55	T	-

T - Time limit of 24 hours exceeded without solution.

5.5. Concluding remarks and future research directions

This study introduces a new version of the off-line admission planning problem under ambiguity distributions of the patient's length of stay. The methodology to solve the MILP problem is based on the dualization of the inner-problem. The approach is suitable for ambiguity-averse decision-makers who choose to make robust decisions, albeit not over-conservative under extreme scenarios. The findings can be used as guidance for practitioners to make robust tactical-operational decisions, such as the case-mix of patients to admit from a waiting list and room scheduling and allocation.

We analyze the benefit of coordinated decisions of scheduling and allocation (i.e., tactical-operational) assuming time-varying bed capacity for the patient-to-room

admission problem under stochastic length of stay. The integrated framework accounts for multi-specialty and several patient types differing by clinical priorities. Through numerical studies employing real data, we have shown that including information not only on the patient's priority but also on the daily capacity availability, together with the uncertain LoS can improve the patient admission. Accordingly, the most priority patients are not the ones that necessarily offer the best system performance; the available bed capacity is also an upper bound of the allocation decisions and cause of bottlenecks for patient flow. Through the analysis of the bed utilization metric, we found that a hospital that is subject to both, uncertainty in the LoS and variable capacity availability will inevitably have under-utilization in some periods while over-utilization in others. Nevertheless, the proposed tactical plan helps to anticipate the admission decisions resulting in better utilization of idle bed capacity.

The other interesting finding is that the admission decisions are highly determined according to the assumptions made over the uncertain LoS, in terms of the solution method and the level of conservatism. We evaluated the reliability of the DRO^{APP} approach by performing an out-of-sample analysis, including ambiguity-averse (i.e., RO^{APP}, DRO^{APP}) and risk-neutral (i.e., SP^{APP}, DT^{APP}) models. We showed that in comparison with the SP^{APP}, RO^{APP}, and DT^{APP} approaches, the DRO^{APP} model allows more flexibility in the decisions regarding possible changes in future scenarios. We believe that for a healthcare institution that aims to protect the patient's well-being, this framework is more desirable. Through a sensitivity analysis, we demonstrate the influence of managing the overstay in the decision process by including the threshold constraint of the maximum days of overstay, i.e., early discharge days. We acknowledge that this tactical policy is prone to cause delays and waiting times at the operational level, but it also helps increase the overall rate of admission and bed utilization. It would be interesting to perform a further study

to evaluate the optimal value of this parameter as a multi-objective measure of capacity availability under uncertain LoS.

Regarding the practical use of the model, we notice that it could be executed in a real hospital environment. Of course, additional developments will be needed for the implementation, like software interfaces and data handling. Typically, a planning model like this might be implemented in a rolling horizon fashion. Rolling horizon applications, as are used in other industries, many times need to fix decisions for the immediate future, in order to avoid excessive rescheduling. This could also be implemented here, and, for example, the first-week schedule could be considered "frozen" for the next iterations. We plan to explore, in the future, the potential practical implementation of the approach in this paper.

The proposed model can also be extended to incorporate additional features, such as gender and age policy, room equipment requirements, and unscheduled patients. The management of waiting times in the admission process is also crucial. Such aspects can be easily included in the current model without modifying the main structure. Additionally, under the proposed framework, future research may focus on the study of upstreams capacities as the ORs planning to account for a more integrated approach of advance scheduling. Further study is needed to investigate larger instances and methods to improve computational efficiency, given the complex combinatorial structure of the problem. Finally, the proposed MILP formulation structure can be broader applied to any problem in which the service time is continuous and can not be interrupted once it started. The modeling scheme can also be generalized to other optimization problems with uncertain service duration, e.g., project scheduling or process planning and scheduling with multiple parallel resources.

Chapter 6. CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this chapter, a summary of the thesis is provided. Section 6.1 presents an overview of the chapters content. The most relevant conclusions of the research conducted in this thesis are outlined in Section 6.2. Finally, Section 6.3 indicates the future research directions.

6.1. Thesis overview

The main objective of this thesis is to model and solve a hierarchical decision-making process, in multiple stages, for admission planning in the healthcare system. By implementing a hierarchical framework, we aim to improve consistency in the decision-making process, by guaranteeing the proper coordination between different temporal levels of decision. The problem of intertemporal consistency arises when making aggregate plans at the tactical level, which are constrained by disaggregated decisions at the operational level. This decision-making process is prevalent in many industries where long-term decisions are taken based on forecasts and under subjective and imprecise conditions which should be fulfilled in the short term that is subject to uncertainty.

We study the admission planning problem for the inpatient service in public hospitals, that is constrained for bed capacity. This problem is subject to several sources of uncertainty, such as patient arrival, LoS, and resource availability, which difficult the admission process, causing delays, unnecessary waiting times, rejections, and even early death of patients. We employed optimization methods under uncertainty in a multi-stage fashion to integrate different decision stages linked in time, aiming to guarantee the expected levels of service and resource utilization. The studies presented in this thesis are based on real data from a public hospital in Chile.

In Chapter 2 we studied hierarchical decision methodologies under uncertainty. We presented an overview of decision-making methodologies that allow modeling intertemporal problems under uncertainty. In particular, we described the basics of the methodologies, two-stage stochastic optimization, robust optimization, and distributionally robust optimization. Besides, a brief review of the up-to-date state of the art of the methodologies mentioned above was provided. The chapter refers to the relevant literature in the field of operations research, which can be used as a guide to understanding the theoretical foundations in decision making under uncertainty.

In Chapter 3 we developed a bi-objective stochastic approach to study the allocation decisions in the admission planning problem, considering an intertemporal approach at the tactical-operational levels. The allocation model aims to balance the service level by considering both hospital and patient perspectives. The intertemporal approach seeks to overcome the infeasibilities that might appear at the operational level due to uncertainty, such as rejections and long waiting times of patients, by anticipating decisions at the tactical level. We proposed a TSO model defined as a mixed-integer linear programming problem, under demand and capacity availability uncertainties, divided into two stages: the first stage variables are reserve capacity decisions for patient groups, and the second-stage variables are decisions about patient allocation and bed utilization deviation. We assumed full knowledge of the probability distributions of the uncertain parameters, obtained from historical data. The purpose of the bi-objective approach is to evaluate the trade-off between the conflicting objectives: resource utilization deviation and service cost.

To derive an equivalent finite-dimensional problem of the TSO model, we implemented a SAA approach, which creates one set of second-stage variables for every possible scenario of the uncertain parameters. The bi-objective problem was solved by

constructing the Pareto efficient curve, which was obtained by employing the weighted-sum method.

Through an extensive numerical study employing real data from a public hospital in Chile, we found that a TSO approach in a bi-objective framework is a suitable methodology for improving consistency at the tactical-operational levels for the allocation problem under bed capacity constraints when enough information is available to describe the uncertainty. This chapter is based on the published paper Batista, Vera, and Pozo (2020).

In Chapter 4 we proposed an alternative framework for modeling service-time-type constraints for problems that cannot be interrupted once allocated. From the mathematical modeling perspective, the proposed framework is related to the problem of appointment on multiple servers with random service duration. It is also linked to the parallel machine problem found in the job shop scheduling literature. Such problems usually rely on integer programming formulations, in which the uncertain parameter is modeled as a summation over a rolling time windows constraint or through integer parameters. We proposed a set of linear constraints for multi-period problems where preemption is not allowed. The proposed MILP formulation considers a single binary variable and continuous service time on the right-hand side of the allocation constraint, enhancing current admission planning (or appointment scheduling) models. Thus, it facilitates the implementation of existing algorithms (e.g., dual-based methods, or bender decomposition) that consider uncertainty, such as stochastic programming, robust optimization, and distributionally robust optimization. We illustrated the applicability of the proposed modeling framework in Chapter 5 and in the published paper Batista, Pozo, and Vera (2020).

In Chapter 5 we studied the robustness of decisions in an intertemporal decision framework for the admission planning problem. The aim was to find a balance between

robustness and consistency under limited distributional information on the patient's length of stay. The consistency question is whether it is possible to achieve a robust tactical plan of admission while guaranteeing a feasible operational plan subject to uncertainty.

We extended the model presented in Chapter 3 to develop a more detailed approach of the APP considering allocation and scheduling decisions under uncertain LoS. This integrated framework is developed to consider coordinated decisions of allocation and scheduling for a multi-specialty, multi-priority, and time-varying capacity framework. The proposed model determines at the tactical level, date and time of admission, along with time allowances and room allocation of patients. These decisions are then constrained at the operational level, where the LoS is uncertain, causing several disruptions in the planning, such as patient overstay.

Due to the limited information available of the patient LoS in hospitals, we adopted a mixed-integer linear DRO approach. It considers that known information is limited only to the first moment and the support set of the true probability distribution of the LoS. The model aims to maximize the net benefit of patient admission, considering the expected cost of overstay represented by an ambiguity-averse expectation measure of the operational cost. To solve the model, we derived a tractable solution methodology employing dual theory. Thus, the infinite-dimensional DRO is reformulated into an exact deterministic equivalent MILP model.

Through an extensive computational study, we illustrated the robustness of the DRO approach by comparing it with alternative frameworks, namely, deterministic, TSO, and RO, employing a real data set from a public hospital in Chile. We also proposed a reliability metric for benchmarking with the different approaches and the conventional cost-based metrics in out-of-sample analysis. The experiments showed that a DRO approach is suitable for ambiguity-averse decision-makers who choose to make robust

decisions, albeit not over-conservative under extreme scenarios. This chapter is based on the paper Batista et al. (2021).

6.2. Conclusions

The general conclusions of the research conducted in this thesis, which support the initial hypothesis proposed in Subsection 1.6.1, are listed below:

1. Extensive numerical analysis and validation employing real data demonstrated that reactive decision-making practices based on deterministic models at a single level of temporal decision are prone to suboptimal solutions for the admission planning problem under bed capacity constraints. Integrated approaches able to capture the operational impact of tactical decisions guarantee better performance of the admission plan.
2. Two-stage stochastic optimization provides a suitable framework to solve the admission planning problem under a tactical-operational intertemporal decision setting. This decision framework in two stages effectively describes the admission decisions while considering the stochastic particularities of the problem and providing proper coordination between temporal decision levels. It allows to anticipate aggregated tactical admission decisions (i.e., reserve bed capacity, admission case mix) acknowledging the randomness at the operational level (i.e., LoS, bed availability, patient arrival) through the recourse function.
3. The novel bi-objective two-stage stochastic approach effectively evaluates conflicting objectives that arise naturally in the healthcare setting. Contrary to current practice in which decisions are taken favoring a single objective, the framework permits to trade-off between hospital and patient perspective

within an hierarchical framework of decision. The best setting of the trade-off evaluation will depend on the goal settle by the decision-maker.

4. The SAA solution methodology employed in the bi-objective two-stage stochastic approach guarantees optimality and efficient computational performance for tactical admission planning. It requires seconds to solve to optimality due to the problem's implicit network structure and the fact that patient demands are integer numbers.
5. Two-stage distributionally robust optimization provides a suitable framework to improve consistency between tactical-operational admission decisions while guaranteeing robustness. The solutions derived for the tactical plan were shown to be highly reliable for extreme scenarios of the uncertain LoS, providing a valid operational plan of allocation.
6. The trade-off between consistency and robustness for the admission planning problem is guaranteed with a two-stage distributionally robust optimization approach. The price of robustness that affects the total benefit of admission is compensated with the diminution of the operational infeasibilities, such as delays, transfers, and overstay, that affect the patient's well-being.
7. A two-stage stochastic formulation with simple recourse can achieve efficient solutions for the admission planning problem under expected operational function values. However, when higher conservatism levels are preferred, and there is limited distributional information of the uncertain parameters, a distributionally robust optimization approach provides the most cost-efficient combination of consistency and robustness, not over conservative as traditional robust optimization.
8. The admission planning problem, although being characterized in general as a combinatorial NP-hard problem, can be solved to optimality in reasonable

computational time (for tactical planning) employing a distributionally robust optimization approach. Exact solutions are obtained through a finite-deterministic mixed-integer linear scenario-based reformulation of the cost function. The enhancement in resolution time is also achieved by employing a modeling framework relying on a single binary variable of allocation and by reducing the combinatorial problem's search space.

9. Managerial insights to practitioners:
   9.1. The admission policy of reserving bed capacities for patient groups at the tactical level (weekly) for later allocation at the operational level (daily), constrained by demand and capacity availability uncertainty, provides better performance, in terms of resource utilization and level of service, than deciding the patient admission daily.
   9.2. Prioritizing deviation in resource utilization over the cost of service guarantees a better balanced patient group allocation but at the expense of higher unmet demand.
   9.3. Prioritizing cost of service over the resource utilization deviation will result in less reserve of internal capacity for prioritized patients groups, when the queue level is constrained. The allocation of those patients in temporary and external beds is the most cost-effective admission policy.
   9.4. For a system subject to demand and capacity availability uncertainty, as lower the fixed target of utilization defined strategically, better flexibility is obtained in the admission process in terms of resource utilization.
   9.5. Deciding the admission of a patient, including information not only on the patient's priority but also on the uncertain LoS and daily capacity availability, provides an overall better performance of the admission plan, than admitting based only on patient priority.

9.6. Determining the patient's time allowances in a system under uncertain LoS is highly dependent on the level of conservatism of the decision-maker. For a risk-neutral decision-maker that considers expected values of the patient's LoS, shorter stay durations are reserved for a high number of patients. On the contrary, for a risk-averse decision-maker, longer stay durations are reserved for a small number of patients. Note that a conservative approach to admission decisions is advisable for a multi-priority system.

9.7. For a system with time-varying bed capacity availability and LoS uncertainty, achieving high bed occupancy targets is a challenge. However, if the tactical admission plan is made within an intertemporal framework of decisions, it is possible to anticipate the operational level's inconsistencies, such as idle capacity, to efficient use of available resources.

9.8. The early discharge policy, when planning the admissions at the tactical level – in a system subject to LoS uncertainty – increases the overall ratio of admission and bed utilization. When considering a conservative approach such as the RO^{APP}, the increment in the number of admissions is not significant. For a less conservative approach such as the DRO^{APP}, more admissions can be planned and still achieve a good balance between robustness and consistency.

6.3. Future research directions

This thesis was developed under the premises of several assumptions that may be relaxed to meet practical requirements in the hospital setting; still, several aspects remain open for future research. This subsection suggests further research directions related to this thesis, outlined below:

1. **To explore a three-level hierarchical decision-making process**.

   We considered an intertemporal decision system of two temporal levels: tactical and operational. We assumed that capacity at the strategic level, i.e., bed capacity is fixed. It would be interesting to consider a three-level framework, including the strategic level at the top, to determine hospital facilities' bed capacity requirements, constrained by random arrivals, LoS, and capacity availability at the tactical-operational level to minimize bed shortages, patient diversion, and waiting time.

2. **To analyze the network structure of the APP under bed capacity constraints**.

   This thesis was developed to improve capacity decisions for a single hospital. However, most healthcare systems in many countries are structured as a network of hospitals serving communities in a specific area. A network framework can be studied in which several hospitals share resources to balance supply and demand under uncertain conditions of patient arrival, LoS, and capacity availability. We believe this cooperative framework can help evaluate capacity decisions from both a temporal and spatial perspective while improving resource utilization and service level.

3. **To study the target of bed utilization as a strategic decision constrained by operational uncertainty**.

   The target of utilization (commonly defined as a fixed performance measure) has been widely criticized (Green, 2002) because it does not capture the variability at the operational level. This measure can be defined as a strategic decision variable to analyze the robustness of the assumption of a fixed 85% target value, thereby determining its optimum value under an intertemporal framework of decisions under uncertainty.

4. **To develop a service level function that fairly measures patient prioritization in the admission process**.

   From an extensive literature review in the context of capacity planning in hospitals, we observed that very few studies are focused on developing effective prioritization methods to improve the admission process. Although there is extensive literature on prioritization techniques for emergency triage (Fernandes et al., 1999), these studies are focused on categorizing patients based on primary symptoms. However, a quality function that considers not only these aspects but a collective categorization, including waiting times and use of resources, is still lacking.

5. **To extend the APP under LoS uncertainty, to include other resources in addition to bed capacity**.

   In this thesis, we assumed the *beds* as a measure of capacity in a hospital, which is a reasonable assumption since, without beds, admission cannot be performed; the beds are where the patients have to be allocated during their entire stay. This assumption considers that other resources, such as special equipment, medical staff, and nurses, are linearly related to the available bed capacity. However, a more integrated admission schedule can be considered

by developing an advanced system under several capacity constraints and intertemporal decisions. For instance, there exist a vast literature on nurse rostering (Burke et al., 2004) and operating room scheduling (Cardoen et al., 2010) that can be incorporated. Insights from such studies can be used to complement our proposed intertemporal decision approach.

6. **To extend the APP under LoS uncertainty to consider larger instances**.
   The APP under uncertain LoS proposed in Chapter 5, consider aggregated decisions of patient-to-room assignment. This assumption by all maters reduced the search space of the model, improving computational performance. A more detailed approach can be further developed to consider more of the complexities of the hospital setting, such as the *patient-to-bed* assignment problem of patients allocated to individual beds, including patient preferences and gender assignment constraints. It would be interesting to further explore the structure of the problem (5.9)–(5.20) in Subsection 5.2.2. This problem has a network structure that can be exploited to improve computational efficiency for larger instances.
7. **To develop enhanced methods to solve a more complex combinatorial APP**.
   As we detailed in Chapter 2, several solution methodologies have been proposed to solve optimization problems under uncertainty. The approaches are based on exact or approximated solution methods, such as sampling methods, affine decision rules, and decomposition-based methods. In order to develop more detailed admission planning problems, such as the patient-to-bed assignment under uncertain LoS, new solution methods will have to be investigated to improve computational efficiency, given the complex combinatorial structure of the problem.

8. **To include additional distributional information for the DRO model under uncertain LoS**.

   The APP under uncertain LoS proposed in Chapter 5 was developed considering first-moment and support information of the uncertain LoS. However, as more information is included in the ambiguity set, the solution tends to be less conservative. The proposed model can be extended to include more distributional information, such as variance and co-variance. The correlation between the patient's stay duration can be considered for patients allocated in the same room. This particular case is useful for isolated care units of contagious diseases.

9. **To explore the use of additional techniques, such as machine learning, in conjunction with optimization for the APP**.

   The APP under uncertainty to improve patient access and resource utilization can also be analyzed in conjunction with other techniques such as prediction models. The output of a prediction model can be used as an input of the optimization model to obtain better accuracy in the decision process. For instance, capacity availability, Length of Stay, and patient discharge can be predicted through artificial intelligent algorithms.

References

Adan, I., Bekkers, J., Dellaert, N., Jeunet, J., & Vissers, J. (2011). Improving operational effectiveness of tactical master plans for emergency and elective patients under stochastic demand and capacitated resources. *European Journal of Operational Research*, *213*(1), 290–308.

Adan, I., Bekkers, J., Dellaert, N., Vissers, J., & Yu, X. (2009). Patient mix optimisation and stochastic resource requirements: A case study in cardiothoracic surgery planning. *Health Care Management Science*, *12*(2), 129.

Adan, I., & Vissers, J. (2002). Patient mix optimisation in hospital admission planning: a case study. *International Journal of Operations & Production Management*, *22*(4), 445–461.

Alvarez, P. P., Espinoza, A., Maturana, S., & Vera, J. (2020). Improving consistency in hierarchical tactical and operational planning using robust optimization. *Computers & Industrial Engineering*, *139*, 106112.

American Hospital Association [AHA]. (2019). *Hospital statistics 2019.* Chicago, Ill.: AHA.

Anthony, R. N. (1965). *Planning and control systems: A framework for analysis*. Division of Research, Graduate School of Business Administration, Harvard University.

Bachouch, R. B., Guinet, A., & Hajri-Gabouj, S. (2012). An integer linear model for hospital bed planning. *International Journal of Production Economics*, *140*(2), 833–843.

Bansal, M., Huang, K.-L., & Mehrotra, S. (2018). Decomposition algorithms for two-stage distributionally robust mixed binary programs. *SIAM Journal on Optimization*, *28*(3), 2360–2383.

Baru, R. A., Cudney, E. A., Guardiola, I. G., Warner, D. L., & Phillips, R. E. (2015). Systematic review of operations research and simulation methods for bed management. In *IIE Annual Conference Proceedings* (p. 298). Institute of Industrial and Systems Engineers (IISE).

Barz, C., & Rajaram, K. (2015). Elective patient admission and scheduling under multiple resource constraints. *Production and Operations Management*, *24*(12), 1907–1930.

Bastos, L. S., Marchesi, J. F., Hamacher, S., & Fleck, J. L. (2019). A mixed integer programming approach to the patient admission scheduling problem. *European Journal of Operational Research*, *273*(3), 831–840.

Batista, A., Pozo, D., & Vera, J. (2020). Stochastic time-of-use-type constraints for uninterruptible services. *IEEE Transactions on Smart Grid*, *11*(1), 229-232.

Batista, A., Pozo, D., & Vera, J. (2021). Managing the unknown: A distributionally robust model for the admission planning problem under uncertain length of stay. *Computers & Industrial Engineering*, 107041. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0360835220307117`

Batista, A., Vera, J., & Pozo, D. (2020). Multi-objective admission planning problem: A two-stage stochastic approach. *Health Care Management Science*, *23*, 51-65.

Beaudoin, D., Frayret, J.-M., & LeBel, L. (2008). Hierarchical forest management with anticipation: an application to tactical–operational planning integration. *Canadian Journal of Forest Research*, *38*(8), 2198–2211.

Bedregal, P., Ferrer, J., Figueroa, B., Tellez, C., Vera, J., & Zurob, C. (2017). La espera en el sistema de salud chileno: una oportunidad para poner a las personas al centro. *Pontifica Universidad Católica de Chile. Recuperado de https://politicaspublicas. uc. cl/wp-content/uploads/2017/12/PDF-TEMAS-DELA-AGENDA-102-. pdf*.

Begen, M. A., & Queyranne, M. (2011). Appointment scheduling with discrete random durations. *Mathematics of Operations Research*, *36*(2), 240–257.

Bekker, R., & Koeleman, P. M. (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, *14*(3), 237.

Beliën, J., & Demeulemeester, E. (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, *176*(2), 1185–1204.

Beliën, J., Demeulemeester, E., & Cardoen, B. (2009). A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling*, *12*(2), 147.

Benders, J. F. (1962). Partitioning procedures for solving mixed-variables programming problems. *Numerische Mathematik*, *4*(1), 1–13.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., & Rennen, G. (2013). Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, *59*(2), 341–357.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust optimization* (Vol. 28). Princeton University Press.

Ben-Tal, A., Goryashko, A., Guslitzer, E., & Nemirovski, A. (2004). Adjustable robust solutions of uncertain linear programs. *Mathematical Programming*, *99*(2), 351–376.

Ben-Tal, A., & Nemirovski, A. (1998). Robust convex optimization. *Mathematics of Operations Research*, *23*(4), 769–805.

Ben-Tal, A., & Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, *25*(1), 1–13.

Berk, E., & Moinzadeh, K. (1998). The impact of discharge decisions on health care quality. *Management Science*, *44*(3), 400–415.

Bertsimas, D., Brown, D. B., & Caramanis, C. (2011). Theory and applications of robust optimization. *SIAM review*, *53*(3), 464–501.

Bertsimas, D., & Caramanis, C. (2010). Finite adaptability in multistage linear optimization. *IEEE Transactions on Automatic Control*, *55*(12), 2751–2766.

Bertsimas, D., Doan, X. V., Natarajan, K., & Teo, C.-P. (2010). Models for minimax stochastic linear optimization problems with risk aversion. *Mathematics of Operations Research*, *35*(3), 580–602.

Bertsimas, D., & Goyal, V. (2010). On the power of robust solutions in two-stage stochastic and adaptive optimization problems. *Mathematics of Operations Research*, *35*(2), 284–305.

Bertsimas, D., Gupta, V., & Kallus, N. (2018a). Data-driven robust optimization. *Mathematical Programming*, *167*(2), 235–292.

Bertsimas, D., Gupta, V., & Kallus, N. (2018b). Robust sample average approximation. *Mathematical Programming*, *171*(1-2), 217–282.

Bertsimas, D., Iancu, D. A., & Parrilo, P. A. (2010). Optimality of affine policies in multistage robust optimization. *Mathematics of Operations Research*, *35*(2), 363–394.

Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044.

Bertsimas, D., Litvinov, E., Sun, X. A., Zhao, J., & Zheng, T. (2012). Adaptive robust optimization for the security constrained unit commitment problem. *IEEE transactions on Power Systems*, *28*(1), 52–63.

Bertsimas, D., Pauphilet, J., Stevens, J., & Tandon, M. (2019). Length-of-stay and mortality prediction for a major hospital through interpretable machine learning. *Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA*.

Bertsimas, D., & Popescu, I. (2005). Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, *15*(3), 780–804.

Bertsimas, D., & Sim, M. (2004). The price of robustness. *Operations Research*, *52*(1), 35–53.

Bertsimas, D., Sim, M., & Zhang, M. (2019). Adaptive distributionally robust optimization. *Management Science*, *65*(2), 604–618.

Billingsley, P. (2013). *Convergence of probability measures*. John Wiley & Sons.

Birge, J. R., & Louveaux, F. (2011). *Introduction to stochastic programming*. Springer Science & Business Media.

Bitran, G. R., & Hax, A. C. (1977). On the design of hierarchical production planning systems. *Decision Sciences*, *8*(1), 28–55.

Brailsford, S., & Vissers, J. (2011). OR in healthcare: A European perspective. *European Journal of Operational Research*, *212*(2), 223–234.

Breton, M., & El Hachem, S. (1995). Algorithms for the solution of stochastic dynamic minimax problems. *Computational Optimization and Applications*, *4*(4), 317–345.

Burke, E. K., De Causmaecker, P., Berghe, G. V., & Van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of Scheduling*, *7*(6), 441–499.

Calafiore, G. C., & El Ghaoui, L. (2006). On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, *130*(1), 1–22.

Cardoen, B., Demeulemeester, E., & Beliën, J. (2010). Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, *201*(3), 921–932.

Carravilla, M. A., & de Sousa, J. P. (1995). Hierarchical production planning in a make-to-order company: A case study. *European Journal of Operational Research*, *86*(1), 43–56.

Carrión, M., & Arroyo, J. M. (2006). A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem. *IEEE Transactions on Power Systems*, *21*(3), 1371–1378.

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, *12*(4), 519–549.

Ceschia, S., & Schaerf, A. (2011). Local search and lower bounds for the patient admission scheduling problem. *Computers & Operations Research*, *38*(10), 1452–1463.

Ceschia, S., & Schaerf, A. (2012). Modeling and solving the dynamic patient admission scheduling problem under uncertainty. *Artificial Intelligence in Medicine*, *56*(3), 199–205.

Ceschia, S., & Schaerf, A. (2016). Dynamic patient admission scheduling with operating room constraints, flexible horizons, and patient delays. *Journal of Scheduling*, *19*(4), 377–389.

Charnes, A., & Cooper, W. W. (1959). Chance-constrained programming. *Management science*, *6*(1), 73–79.

Chen, X., Sim, M., & Sun, P. (2007). A robust optimization perspective on stochastic programming. *Operations Research*, *55*(6), 1058–1071.

Conejo, A. J., Carrión, M., & Morales, J. M. (2010). *Decision making under uncertainty in electricity markets* (Vol. 1). Springer.

Conejo, A. J., Castillo, E., Minguez, R., & Garcia-Bertrand, R. (2006). *Decomposition techniques in mathematical programming: engineering and science applications*. Springer Science & Business Media.

Conforti, D., Guerriero, F., Guido, R., Cerinic, M. M., & Conforti, M. L. (2011). An optimal decision making model for supporting week hospital management. *Health Care Management Science*, *14*(1), 74–88.

Dantzig, G. B. (1955). Linear programming under uncertainty. *Management Science*, *1*(3-4), 197–206.

Deb, K. (2014). Multi-objective optimization. In *Search methodologies* (pp. 403–449). Springer.

Delage, E., & Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, *58*(3), 595–612.

Demeester, P., Souffriau, W., De Causmaecker, P., & Berghe, G. V. (2010). A hybrid tabu search algorithm for automatically assigning patients to beds. *Artificial Intelligence in Medicine*, *48*(1), 61–70.

Dempster, M. A. H., Fisher, M., Jansen, L., Lageweg, B., Lenstra, J. K., & Rinnooy Kan, A. (1981). Analytical evaluation of hierarchical planning systems. *Operations Research*, *29*(4), 707–716.

Deng, Y., & Shen, S. (2016). Decomposition algorithms for optimizing multi-server appointment scheduling with chance constraints. *Mathematical Programming*, *157*(1), 245–276.

Deng, Y., Shen, S., & Denton, B. (2019). Chance-constrained surgery planning under conditions of limited and ambiguous data. *INFORMS Journal on Computing*.

Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, *35*(11), 1003–1016.

Denton, B. T., Miller, A. J., Balasubramanian, H. J., & Huschka, T. R. (2010). Optimal allocation of surgery blocks to operating rooms under uncertainty. *Operations Research*, *58*(4-part-1), 802–816.

Dobson, G., Lee, H.-H., & Pinker, E. (2010). A model of ICU bumping. *Operations Research*, *58*(6), 1564–1576.

Dupačová, J. (1987). The minimax approach to stochastic programming and an illustrative application. *Stochastics: An International Journal of Probability and Stochastic Processes*, *20*(1), 73–88.

Durán, G., Rey, P. A., & Wolff, P. (2017). Solving the operating room scheduling problem with prioritized lists of patients. *Annals of Operations Research*, *258*(2), 395–414.

El Ghaoui, L., & Lebret, H. (1997). Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, *18*(4), 1035–1064.

El Ghaoui, L., Oustry, F., & Lebret, H. (1998). Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, *9*(1), 33–52.

Erdoğan, E., & Iyengar, G. (2006). Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, *107*(1-2), 37–61.

Esfahani, P. M., & Kuhn, D. (2018). Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, *171*(1-2), 115–166.

Fanjiang, G., Grossman, J. H., Compton, W. D., & Reid, P. P. (2005). *Building a better delivery system: a new engineering/health care partnership*. National Academies Press.

Fernandes, C. M., Wuerz, R., Clark, S., Djurdjev, O., & Group, M. O. R. (1999). How reliable is emergency department triage? *Annals of Emergency Medicine*, *34*(2), 141–147.

Gabrel, V., Murat, C., & Thiele, A. (2014). Recent advances in robust optimization: An overview. *European Journal of Operational Research*, *235*(3), 471–483.

Gallivan, S., Utley, M., Treasure, T., & Valencia, O. (2002). Booked inpatient admissions and hospital capacity: mathematical modelling study. *Bmj*, *324*(7332), 280–282.

Gao, R., & Kleywegt, A. J. (2016). Distributionally robust stochastic optimization with Wasserstein distance. *arXiv preprint arXiv:1604.02199*.

Gemmel, P., & Van Dierdonck, R. (1999). Admission scheduling in acute care hospitals: does the practice fit with the theory? *International Journal of Operations & Production Management*, *19*(9), 863–878.

Ghaoui, L. E., Oks, M., & Oustry, F. (2003). Worst-case value-at-risk and robust portfolio optimization: A conic programming approach. *Operations Research*, *51*(4), 543–556.

Goh, J., & Sim, M. (2010). Distributionally robust optimization and its tractable approximations. *Operations Research*, *58*(4-part-1), 902–917.

González Cobos, N. (2019). *Robust generation scheduling in electricity markets* (Unpublished doctoral dissertation). Universidad de Castilla-La Mancha.

Gorissen, B. L., Yanıkoğlu, İ., & den Hertog, D. (2015). A practical guide to robust optimization. *Omega*, *53*, 124–137.

Green, L. V. (2002). How many hospital beds? *Inquiry: The Journal of Health Care Organization, Provision, and Financing*, *39*(4), 400–412.

Green, L. V. (2005). Capacity planning and management in hospitals. In *Operations Research and Health Care* (pp. 15–41). Springer.

Green, L. V. (2012). The vital role of operations analysis in improving healthcare delivery. *Manufacturing & Service Operations Management*, *14*(4), 488–494.

Green, L. V., & Nguyen, V. (2001). Strategies for cutting hospital beds: the impact on patient service. *Health Services Research*, *36*(2), 421.

Groot, P. M. A. (1993). *Decision support for admission planning under multiple resource constraints*. Technische Universiteit Eindhoven.

Guido, R., Groccia, M. C., & Conforti, D. (2018). An efficient matheuristic for offline patient-to-bed assignment problems. *European Journal of Operational Research*, *268*(2), 486–503.

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

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*(9), 800–819.

Hall, R. W. (2012). *Handbook of healthcare system scheduling*. Springer.

Hanasusanto, G. A., Roitch, V., Kuhn, D., & Wiesemann, W. (2015). A distributionally robust perspective on uncertainty quantification and chance constrained programming. *Mathematical Programming*, *151*(1), 35–62.

Hans, E. W., Van Houdenhoven, M., & Hulshof, P. J. (2012). A framework for healthcare planning and control. In *Handbook of healthcare system scheduling* (pp. 303–320). Springer.

Harper, P. R., & Shahani, A. (2002). Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, *53*(1), 11–18.

Hartmann, S., & Briskorn, D. (2010). A survey of variants and extensions of the resource-constrained project scheduling problem. *European Journal of Operational Research*, *207*(1), 1–14.

Hax, A., & Meal, H. (1975). Hierarchical integration of production planning and scheduling; studies in management sciences. *Logistics. New York: North Holland-American Elsevier*.

He, L., Madathil, S. C., Oberoi, A., Servis, G., & Khasawneh, M. T. (2019). A systematic review of research design and modeling techniques in inpatient bed management. *Computers & Industrial Engineering*, *127*, 451–466.

Heitsch, H., & Römisch, W. (2003). Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications*, *24*(2-3), 187–206.

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

Higle, J. L., & Sen, S. (1991). Stochastic decomposition: An algorithm for two-stage linear programs with recourse. *Mathematics of Operations Research*, *16*(3), 650–669.

Homem-de Mello, T., & Bayraksan, G. (2014). Monte carlo sampling-based methods for stochastic optimization. *Surveys in Operations Research and Management Science*, *19*(1), 56–85.

Hu, Z., & Hong, L. J. (2013). Kullback-Leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*.

Hu, Z., Hong, L. J., & So, A. M.-C. (2013). Ambiguous probabilistic programs. *Available at Optimization Online*.

Hulshof, P. J., Boucherie, R. J., Hans, E. W., & Hurink, J. L. (2013). Tactical resource allocation and elective patient admission planning in care processes. *Health Care Management Science*, *16*(2), 152–166.

Hulshof, P. J., Kortbeek, N., Boucherie, R. J., Hans, E. W., & Bakker, P. J. (2012). Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health Systems*, *1*(2), 129–175.

Hulshof, P. J., Mes, M. R., Boucherie, R. J., & Hans, E. W. (2016). Patient admission planning using approximate dynamic programming. *Flexible Services and Manufacturing Journal*, *28*(1-2), 30–61.

Jebali, A., & Diabat, A. (2015). A stochastic model for operating room planning under capacity constraints. *International Journal of Production Research*, *53*(24), 7252–7270.

Jebali, A., & Diabat, A. (2017). A chance-constrained operating room planning with elective and emergency cases under downstream capacity constraints. *Computers & Industrial Engineering*, *114*, 329–344.

Jiang, R., & Guan, Y. (2016). Data-driven chance constrained stochastic program. *Mathematical Programming*, *158*(1-2), 291–327.

Jiang, R., Shen, S., & Zhang, Y. (2017). Integer programming approaches for appointment scheduling with random no-shows and service durations. *Operations Research*, *65*(6), 1638–1656.

Joosten, T., Bongers, I., & Janssen, R. (2009). Application of lean thinking to health care: issues and observations. *International Journal for Quality in Health Care*, *21*(5), 341–347.

Kantorovich, L. V., & Rubinstein, G. S. (1958). On a space of completely additive functions. *Vestnik Leningrad. Univ*, *13*(7), 52–59.

Kira, D., Kusy, M., & Rakita, I. (1997). A stochastic linear programming approach to hierarchical production planning. *Journal of the Operational Research Society*, *48*(2), 207–211.

Kleywegt, A. J., Shapiro, A., & Homem-de Mello, T. (2002). The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, *12*(2), 479–502.

Kokangul, A. (2008). A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit. *Computer Methods and Programs in Biomedicine*, *90*(1), 56–65.

Kong, Q., Lee, C.-Y., Teo, C.-P., & Zheng, Z. (2013). Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations Research*, *61*(3), 711–726.

Korf, R. E. (2002). A new algorithm for optimal bin packing. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence* (pp. 731–736).

Kortbeek, N., Braaksma, A., Burger, C. A., Bakker, P. J., & Boucherie, R. J. (2015). Flexible nurse staffing based on hourly bed census predictions. *International Journal of Production Economics*, *161*, 167–180.

Kuhn, D., Wiesemann, W., & Georghiou, A. (2011). Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming*, *130*(1), 177–209.

Kusters, R. J., & Groot, P. M. (1996). Modelling resource availability in general hospitals design and implementation of a decision support model. *European Journal of Operational Research*, *88*(3), 428–445.

Laguna, J., Arbeloa, G., Arbeloa, P., Oyarzabal, M., & Osakidetza. (2000). *GRD: manual de descripción de los grupos relacionados por el diagnóstico (AP-GRD v. 14.1)*. Osakidetza = Servicio Vasco de Salud. Retrieved from `https://books.google.ru/books?id=kruSYgEACAAJ`

Lakshmi, C., & Iyer, S. A. (2013). Application of queueing theory in health care: A literature review. *Operations research for health care*, *2*(1-2), 25–39.

Landau, H. J. (1987). *Moments in mathematics* (Vol. 37). American Mathematical Society.

Li, X., Liu, D., Geng, N., & Xie, X. (2018). Optimal ICU admission control with premature discharge. *IEEE Transactions on Automation Science and Engineering*.

Liu, N., Truong, V.-A., Wang, X., & Anderson, B. R. (2019). Integrated scheduling and capacity planning with considerations for patients' length-of-stays. *Production and Operations Management*, *28*(7), 1735–1756.

Lobos, A., & Vera, J. R. (2016). Intertemporal stochastic sawmill planning: Modeling and managerial insights. *Computers & Industrial Engineering*, *95*, 53–63.

Mak, H.-Y., Rong, Y., & Zhang, J. (2014). Appointment scheduling with limited distributional information. *Management Science*, *61*(2), 316–334.

Mazier, A., Xie, X., & Sarazin, M. (2010). Scheduling inpatient admission under high demand of emergency patients. In *2010 IEEE International Conference on Automation Science and Engineering* (pp. 792–797). IEEE.

Mcdiarmid, C., Habib, M., Ramirez-Alfonsin, J., & Reed, B. (1998). Probabilistic methods for algorithmic discrete mathematics. *Algorithms and Combinatorics Series*, *16*(1-46), 11.

Mehrotra, S., & Papp, D. (2014). A cutting surface algorithm for semi-infinite convex programming with an application to moment robust optimization. *SIAM Journal on Optimization*, *24*(4), 1670–1697.

Mehrotra, S., & Zhang, H. (2014). Models and algorithms for distributionally robust least squares problems. *Mathematical Programming*, *146*(1-2), 123–141.

Meng, F., Qi, J., Zhang, M., Ang, J., Chu, S., & Sim, M. (2015). A robust optimization model for managing elective admission in a public hospital. *Operations Research*, *63*(6), 1452–1467.

Miller, B. L., & Wagner, H. M. (1965). Chance constrained programming with joint constraints. *Operations Research*, *13*(6), 930–945.

Min, D., & Yih, Y. (2010a). An elective surgery scheduling problem considering patient priority. *Computers & Operations Research*, *37*(6), 1091–1099.

Min, D., & Yih, Y. (2010b). Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, *206*(3), 642–652.

Minoux, M. (2012). Two-stage robust LP with ellipsoidal right-hand side uncertainty is NP-hard. *Optimization Letters*, *6*(7), 1463–1475.

Mittal, S., Schulz, A. S., & Stiller, S. (2014). Robust appointment scheduling. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2014).* Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Mokotoff, E. (2001). Parallel machine scheduling problems: A survey. *Asia-Pacific Journal of Operational Research*, *18*(2), 193.

Mondschein, S. V., & Weintraub, G. Y. (2003). Appointment policies in service operations: A critical analysis of the economic framework. *Production and Operations Management*, *12*(2), 266–286.

Mullen, P. M. (2003). Prioritising waiting lists: how and why? *European Journal of Operational Research*, *150*(1), 32–45.

Nam, S.-j., & Logendran, R. (1992). Aggregate production planning—a survey of models and methodologies. *European Journal of Operational Research*, *61*(3), 255–272.

Nguyen, J.-M., Six, P., Antonioli, D., Glemain, P., Potel, G., Lombrail, P., & Le Beux, P. (2005). A simple method to optimize hospital beds capacity. *International Journal of Medical Informatics*, *74*(1), 39–49.

OECD. (2020). *Health spending (indicator).* `https://data.oecd.org/healthres/health-spending.htm`. (doi: 10.1787/8643de7e-en. Accessed: January 8, 2020)

Ostrowski, J., Anjos, M. F., & Vannelli, A. (2012). Tight mixed integer linear programming formulations for the unit commitment problem. *IEEE Transactions on Power Systems*, *27*(1), 39–46.

Ozturk, U. A., Mazumdar, M., & Norman, B. A. (2004). A solution to the stochastic unit commitment problem using chance constrained programming. *IEEE Transactions on Power Systems*, *19*(3), 1589–1598.

Peters-Groot, P. (1993). *Decision support for admission planning under multiple resource constraints* (Unpublished doctoral dissertation). Technische Universiteit Eindhoven.

Pflug, G., & Wozabal, D. (2007). Ambiguity in portfolio selection. *Quantitative Finance*, *7*(4), 435–442.

Pierskalla, W. P., & Brailer, D. J. (1994). Applications of operations research in health care delivery. *Handbooks in operations research and management science*, *6*, 469–505.

Pinedo, M. (2012). *Scheduling* (Vol. 5). Springer.

Popescu, I. (2007). Robust mean-covariance solutions for stochastic optimization. *Operations Research*, *55*(1), 98–112.

Postek, K., den Hertog, D., & Melenberg, B. (2016). Computationally tractable counterparts of distributionally robust constraints on risk measures. *SIAM Review*, *58*(4), 603–650.

Pozo, D., Street, A., & Velloso, A. (2018). An ambiguity-averse model for planning the transmission grid under uncertainty on renewable distributed generation. In *2018 Power Systems Computation Conference (PSCC)* (pp. 1–7). IEEE.

Prékopa, A. (2013). *Stochastic programming* (Vol. 324). Springer Science & Business Media.

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, *1*(1), 18.

Range, T. M., Lusby, R. M., & Larsen, J. (2014). A column generation approach for solving the patient admission scheduling problem. *European Journal of Operational Research*, *235*(1), 252–264.

Rauner, M. S., & Vissers, J. M. (2003). Or applied to health services: Planning for the future with scarce resources. *European Journal of Operational Research*, *150*(1), 1–2.

Rebolledo, R. (2019). Las crueles cifras de las listas de espera en la salud publica: 9.724 fallecidos. *La Izquierda Diario*. Retrieved December 16, 2019, from `http://www.laizquierdadiario.cl/Las-crueles-cifras-de-las-Listas-de-Espera-en-la-Salud-Publica`.

Reid, P., & Grossman, J. (2005). A framework for a systems approach to health care delivery. *Building a better delivery system: A new engineering/health care partnership. National Academies Press, Washington, DC, États-Unis*.

Rosko, M. D. (1988). DRGs and severity of illness measures: An analysis of patient classification systems. *Journal of Medical Systems*, *12*(4), 257–274.

Samiedaluie, S., Kucukyazici, B., Verter, V., & Zhang, D. (2017). Managing patient admissions in a neurology ward. *Operations Research*, *65*(3), 635–656.

Samudra, M., Van Riet, C., Demeulemeester, E., Cardoen, B., Vansteenkiste, N., & Rademakers, F. E. (2016). Scheduling operating rooms: achievements, challenges and pitfalls. *Journal of Scheduling*, *19*(5), 493–525.

Scarf, H. (1958). A min-max solution of an inventory problem. *Studies in the Mathematical Theory of Inventory and Production*, 201–209.

Seung-Chul, K., & Ira, H. (2000). Flexible bed allocation and performance in the intensive care unit. *Journal of Operations Management*, *18*(4), 427–443.

Shang, C., & You, F. (2018). Distributionally robust optimization for planning and scheduling under uncertainty. *Computers & Chemical Engineering*, *110*, 53–68.

Shapiro, A., & Ahmed, S. (2004). On a class of minimax stochastic programs. *SIAM Journal on Optimization*, *14*(4), 1237–1249.

Shapiro, A., Dentcheva, D., & Ruszczyński, A. (2014). *Lectures on stochastic programming: modeling and theory*. SIAM.

Soyster, A. L. (1973). Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, *21*(5), 1154–1157.

Sözüer, S., & Thiele, A. C. (2016). The state of robust optimization. In *Robustness Analysis in Decision Aiding, Optimization, and Analytics* (pp. 89–112). Springer.

Tamiz, M., Mirrazavi, S., & Jones, D. (1999). Extensions of pareto efficiency analysis to integer goal programming. *Omega*, *27*(2), 179–188.

Teixeira, A., & De Oliveira, M. (2015). Operations research on hospital admission systems: a first overview of the 2005-2014 decade. In *Journal of Physics: Conference Series* (Vol. 616, p. 012009). IOP Publishing.

Testi, A., & Tànfani, E. (2009). Tactical and operational decisions for operating room planning: Efficiency and welfare implications. *Health Care Management Science*, *12*(4), 363.

Testi, A., Tanfani, E., Valente, R., Ansaldo, G., & Torre, G. (2008). Prioritizing surgical waiting lists. *Journal of Evaluation in Clinical Practice*, *14*(1), 59–64.

Thompson, S., Nunez, M., Garfinkel, R., & Dean, M. D. (2009). Or practice—efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research*, *57*(2), 261–273.

Turhan, A. M., & Bilgen, B. (2017). Mixed integer programming based heuristics for the patient admission scheduling problem. *Computers & Operations Research*, *80*, 38–49.

Utley, M., Gallivan, S., Davis, K., Daniel, P., Reeves, P., & Worrall, J. (2003). Estimating bed requirements for an intermediate care facility. *European Journal of Operational Research*, *150*(1), 92–100.

Utley, M., Jit, M., & Gallivan, S. (2008). Restructuring routine elective services to reduce overall capacity requirements within a local health economy. *Health Care Management Science*, *11*(3), 240–247.

Vancroonenburg, W., De Causmaecker, P., & Berghe, G. V. (2016). A study of decision support models for online patient-to-room assignment planning. *Annals of Operations Research*, *239*(1), 253–271.

Vancroonenburg, W., De Causmaecker, P., & Berghe, G. V. (2019). Chance-constrained admission scheduling of elective surgical patients in a dynamic, uncertain setting. *Operations Research for Health Care*, *22*, 100196.

Velloso, A., Pozo, D., & Street, A. (2020). Distributionally robust transmission expansion planning: a multi-scale uncertainty approach. *IEEE Transactions on Power Systems*, *35*(5), 3353–3365.

Vijayakumar, B., Parikh, P. J., Scott, R., Barnes, A., & Gallimore, J. (2013). A dual bin-packing approach to scheduling surgical cases at a publicly-funded hospital. *European Journal of Operational Research*, *224*(3), 583–591.

Vissers, J., Adan, I. J., & Bekkers, J. A. (2005). Patient mix optimization in tactical cardiothoracic surgery planning: a case study. *IMA journal of Management Mathematics*, *16*(3), 281–304.

Vissers, J. M., Adan, I. J., & Dellaert, N. P. (2007). Developing a platform for comparison of hospital admission systems: An illustration. *European Journal of Operational Research*, *180*(3), 1290–1301.

Vissers, J. M., Bertrand, J., & De Vries, G. (2001). A framework for production control in health care organizations. *Production Planning & Control*, *12*(6), 591–604.

Wang, S., Li, J., & Peng, C. (2017). Distributionally robust chance-constrained program surgery planning with downstream resource. In *2017 International Conference on Service Systems and Service Management (ICSSM)* (pp. 1–6). IEEE.

Wang, X., Truong, V.-A., & Bank, D. (2018). Online advance admission scheduling for services with customer preferences. *arXiv preprint arXiv:1805.10412*.

Wang, Y., Zhang, Y., & Tang, J. (2019). A distributionally robust optimization approach for surgery block allocation. *European Journal of Operational Research*, *273*(2), 740–753.

Wiesemann, W., Kuhn, D., & Sim, M. (2014). Distributionally robust convex optimization. *Operations Research*, *62*(6), 1358–1376.

Wozabal, D. (2012). A framework for optimization under ambiguity. *Annals of Operations Research*, *193*(1), 21–47.

Yan, H.-S. (2000). Hierarchical stochastic production planning for flexible automation workshops. *Computers & Industrial Engineering*, *38*(4), 435–455.

Yan, H.-S., Xia, Q.-F., Zhu, M.-R., Liu, X.-L., & Guo, Z.-M. (2003). Integrated production planning and scheduling on automobile assembly lines. *IIE Transactions*, *35*(8), 711–725.

Young, J. P. (1965). Stabilization of inpatient bed occupancy through control of admissions. *Hospitals*, *39*(19), 41.

Yue, J., Chen, B., & Wang, M.-C. (2006). Expected value of distribution information for the newsvendor problem. *Operations Research*, *54*(6), 1128–1136.

Žáčková, J. (1966). On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky*, *91*(4), 423–430.

Zeng, B., & Zhao, L. (2013). Solving two-stage robust optimization problems using a column-and-constraint generation method. *Operations Research Letters*, *41*(5), 457–461.

Zhang, Y., Puterman, M. L., Nelson, M., & Atkins, D. (2012). A simulation optimization approach to long-term care capacity planning. *Operations Research*, *60*(2), 249–261.

Zhang, Y., Shen, S., & Erdogan, S. A. (2017). Distributionally robust appointment scheduling with moment-based ambiguity set. *Operations Research Letters*, *45*(2), 139–144.

Zhang, Y., Shen, S., & Erdogan, S. A. (2018). Solving 0–1 semidefinite programs for distributionally robust allocation of surgery blocks. *Optimization Letters*, *12*(7), 1503–1521.

Zhao, L., & Zeng, B. (2012a). An exact algorithm for two-stage robust optimization with mixed integer recourse problems. *submitted, available on Optimization-Online. org*.

Zhao, L., & Zeng, B. (2012b). Robust unit commitment problem with demand response and wind energy. In *2012 IEEE Power and Energy Society general meeting* (pp. 1–8). IEEE.

Zhu, S., & Fukushima, M. (2009). Worst-case conditional value-at-risk with application to robust portfolio management. *Operations Research*, *57*(5), 1155–1168.

Zymler, S. (2010). *Distributionally robust optimization with applications to risk management* (Unpublished doctoral dissertation). Imperial College London.

APPENDIX A. DESIGN OF DATA COLLECTION SHEET OF HOSPITAL DAILY OPERATION

This section introduces the data collection sheet designed in conjunction with the hospital under study to collect data about daily bed capacity requirements and availability[1]. The process of data collection is described in Section A.1 and the spreadsheet design in Section A.2.

A.1. Data collection process in the hospital Central Admission Department

The hospital under study has an EHR system that collects data about the patient's date of admission and discharge, admission care unit, length of stay, diagnosis, DRG weight, and severity index. However, at the time of this research, information about daily bed availability and requirements was not collected. Since this information is crucial for this thesis, in conjunction with the hospital CAD managers, we designed a spreadsheet to record the data to be used as input to the mathematical optimization models and help the hospital administer their resources better.

The data collection process in the hospital under study is currently manual. Daily, the CAD collects data about the status of the different care units (e.g., emergency, medicine, neurology, surgery), including the number of patients on the waiting list, bed availability, and requirements. This data is then included in the spreadsheet.

A sample of the spreadsheet is shown in Figures A.1–A.6, which indicates different modules of the data collection process. For this thesis, we employed data related to bed requirements and availability for the different care services during the period 2017-2019. The collected data was also used to validate the optimization models. Worthy of note is that we acknowledge that a process that is done manually is prone to misleading. However,

[1]Note that for privacy reasons we have omitted the details about the hospital under study.

the use of this spreadsheet to record relevant information of the hospital operations is a first step to build an automated system, especially nowadays, when having relevant data available can help make better decisions.

In the following section, we detail the information contained in each of the spreadsheet modules of the Figures A.1–A.6.

A.2. The data collection spreadsheet design

The first module is presented in Figure A.1, in which data about the number of *hospitalized* patients for different care units is recorded.

As shown in Figure A.1, the data collection spreadsheet contains in its first row the days of the month, differentiating between data collected during the morning (AM) and the afternoon (PM)[2]. At the bottom of each module, the data is summarized. The first column indicates the name of the module. The two last columns calculate the average of the collected information.

MONTH/DAY	1		2		3		4		5		6		7		Average	Average
MAY '2017'	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
HOSPITALIZED - SERVICES																
MEDICINE																
NEUROLOGY																
SURGERY																
TRAUMATOLOGY																
UPC (UCI- UTIM -UCE)																
ACUTE																
TOTAL HOSPITALIZED - SERVICES																

FIGURE A.1. Spreadsheet module: hospitalized.

Figure A.2 contains information about the emergency service, which is important to get an idea of the overflow of patients that will have to be allocated in the internal hospital

[2]The headline of the spreadsheet is repeated in the next figures.

beds. Two modules are indicated, *waiting time for bed* and *status*. The first section is completed with data about the number of patients waiting in the emergency service considering different threshold times (e.g., 8–11, 12-23). The second section indicates the status of the emergency service in terms of the number of patients hospitalized in *triage*[3], the patients who have been discharged, and those waiting in temporary units to be transferred to the internal hospital beds. Finally, the spreadsheet also records data about the number of patients who have died in the emergency service.

MONTH/DAY	1		2		3		4		5		6		7		Average	Average
MAY '2017'	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
WAITING TIME FOR BED - EMERGENCY																
LESS THAN 8 HOURS																
8 - 11 HOURS																
12 - 23 HOURS																
GREATER THAN 24 HOURS																
STATUS - EMERGENCY																
HOSPITALIZED IN TRIAGE																
DISCHARGE EMERGENCY																
AMBULANCES WAITING																
WAITING FOR A CHAIR																
WAITING FOR STRETCHER																
WAITING FOR ADMISSION EMERGENCY																
C2																
C3																
C4																
DEATH IN EMERGENCY																

FIGURE A.2. Spreadsheet modules: waiting time, status, death (emergency service).

Figure A.3 contains the *free beds* module, which is completed with data about the number of available beds in the different care units. We list some of the care units in the hospital, namely, medicine, neurology, and surgery. Besides, data about the number of available beds in the critical patient unit (UPC) is also recorded, differentiating between

[3]Triage is a classification system employed in emergency units to categorize patients according to priorities, e.g., C2, C3, C4.

beds from the intensive care unit (UCI), intermediate care unit (UTIM), and special care unit (UCE).

MONTH/DAY	1		2		3		4		5		6		7		Average	Average
MAY '2017'	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
FREE BEDS 8 AM																
MEDICINE																
NEUROLOGY																
SURGERY																
TOTAL FREE GENERAL BEDS																
UPC																
UCI																
UTIM																
UCE																
TOTAL FREE UPC BEDS																

FIGURE A.3. Spreadsheet module: free beds.

Figures A.4 and A.5 contain the module of *bed requirements*, which is completed with data about the demand for beds (e.g., general, UPC, acute), from the different care units. The requirements will then be, satisfied or not depending on bed availability.

MONTH/DAY	1		2		3		4		5		6		7		Average	Average
MAY '2017'	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
BED REQUIREMENTS																
EMERGENCY																
GENERAL																
UPC																
ACUTE																
TOTAL																
RECOVERY SERVICE																
GENERAL																
UPC																
TOTAL																
UPC																
GENERAL																
UCI																
UTIM																
UCE																
TOTAL																

FIGURE A.4. Spreadsheet module: bed requirements (a).

MONTH/DAY	1		2		3		4		5		6		7		Average	Average
MAY '2017'	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
BED REQUIREMENTS																
EMERGENCY																
GENERAL																
UPC																
ACUTE																
TOTAL																
RECOVERY SERVICE																
GENERAL																
UPC																
TOTAL																
UPC																
GENERAL																
UCI																
UTIM																
UCE																
TOTAL																

FIGURE A.5. Spreadsheet module: bed requirements (b).

Figure A.6 contains the module of *discharge*, which is completed with data about the number of patients discharged from different care units. A distinction is made between the information received in the morning (AM) and the confirmed information in the afternoon (PM). Besides, it is recorded data about the *diverted* patients to the network of hospitals, distinguishing between morning and afternoon executions. Finally, the spreadsheet is completed with data about the number of *blocked beds* in different care units. This information is essential to visualized bed availability over time.

MONTH/DAY	1		2		3		4		5		6		7		Average	Average
MAY '2017'	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM	AM	PM
DISCHARGE ANNOUNCED (AM)																
DISCHARGE CONFIRMED (PM)																
MEDICINE																
SURGERY																
NEUROLOGY																
UPC																
TOTAL																
DIVERTED PATIENTS (NETWORK)																
BEFORE 11 AM																
BEFORE 7 PM																
TOTAL																
BLOCKED BEDS																
GENERAL																
PSYCHIATRY																
UCI																
UTIM																
UTIQ																
UHI																
TOTAL																

FIGURE A.6. Spreadsheet modules: discharge, diverted, blocked beds.

The spreadsheet is currently being used in the hospital to collect relevant data for the admission planning process. Future improvements are still needed to automate the tool, to allow online update in connection with the central EHR.

APPENDIX B. DATA DESCRIPTION OF THE ADMISSION PLANNING PROBLEM APPROACH UNDER UNCERTAIN LENGTH OF STAY

This section details the data input used to solve the APP approach under uncertain LoS, presented in Chapter 5. The dataset was obtained from the EHR of a public hospital in Chile[1]. It contains historical data of patient admission from 2010-2016. The size of the samples differs according to the patient type and room of allocation.

B.1. Statistical analysis of the patient DRG and Severity index

Table B.1 reports the results of the statistical analysis to determine the correlation between the patient DRG and severity index. The study was performed employing a parametric test of Pearson correlation over the available data per patient diagnosis. The samples of patient types are reported in the table footer. The results indicate that the values are positively correlated, with a $P < 0.05$. We difference between high correlation for $\rho \geq 0.5$ and moderate correlation when $\rho < 0.5$.

TABLE B.1. Pearson's correlation for the patient DRG and Severity Illness indexes.

		Severity Illness Index											
		C	F	H	I	K	S	Z	E	J	M	N	T
DRG	C	0.584*	-	-	-	-	-	-	-	-	-	-	-
	F	-	0.522*	-	-	-	-	-	-	-	-	-	-
	H	-	-	0.547*	-	-	-	-	-	-	-	-	-
	I	-	-	-	0.399**	-	-	-	-	-	-	-	-
	K	-	-	-	-	0.453**	-	-	-	-	-	-	-
	S	-	-	-	-	-	0.539*	-	-	-	-	-	-
	Z	-	-	-	-	-	-	0.764*	-	-	-	-	-
	E	-	-	-	-	-	-	-	0.539*	-	-	-	-
	J	-	-	-	-	-	-	-	-	0.343**	-	-	-
	M	-	-	-	-	-	-	-	-	-	0.371**	-	-
	N	-	-	-	-	-	-	-	-	-	-	0.322**	-
	T	-	-	-	-	-	-	-	-	-	-	-	0.493**

Sample of patient type: C = 6299; F = 2090; H = 2421; I = 9514; K = 10722; S = 7464; Z = 1670; E = 3338; J = 3455; M = 2717; N = 4577; T = 3673 - **Correlation:** *High correlation $\rho \geq 0.5$, **Moderate correlation, $\rho < 0.5$

[1]Specific details about the hospital under study have been omitted for the sake of privacy.

B.2. Demand proportion, mean and support of patient type and care unit

Table B.2 describes the distributional information about the patient types and historical proportions of admission. Columns 1 and 2, include the patient's type and related ID, which corresponds to the DRG classification of the patient diagnosis. Each letter is associated with a diagnosis group reported in the EHR of the hospital under study. For a description of each diagnosis-group, the reader is referred to (Laguna et al., 2000). Columns 3 and 4 report the mean vector, $\bar{\boldsymbol{\xi}}$, and support set, Ξ, of the patients. Columns 5–11 indicate the demand proportions of patient's admission to the different rooms.

TABLE B.2. Mean value (in days), support set (in days) and demand proportion per care unit from period 2010–2016.

		Length of Stay data		Admission request by room (%)									
Diagnosis ID	Diagnosis	$\bar{\xi}$	Ξ	Medicine (%)	Surgery (%)	Ophthalmology (%)	Adult transplant (%)	Psychiatry (%)	Adult pensioner (%)	ICU (%)	CMA (%)	Intermediate care Surgery (%)	Intermediate care Medicine (%)
1	C	9.93	[2,26]	7.58	12.89	0.73	51.41	0.00	9.05	10.20	6.20	25.55	4.99
2	F	13.02	[2,27]	1.13	0.03	0.00	0.00	87.67	0.18	0.33	0.00	0.00	0.00
3	H	5.06	[1,15]	0.62	0.41	43.03	0.00	0.00	0.62	0.00	0.00	0.00	0.00
4	I	11.26	[1,26]	29.04	6.33	0.02	0.95	0.05	7.83	29.02	8.86	21.03	51.07
5	K	6.71	[1,21]	8.15	22.11	0.02	2.31	0.00	20.66	21.41	41.83	18.44	8.19
6	S	9.18	[1,25]	2.96	21.87	48.46	0.00	0.22	5.57	11.59	16.47	7.33	2.93
7	Z	5.26	[0, 12]	3.35	1.07	1.69	26.26	0.00	0.33	0.35	3.13	1.59	1.40
8	E	8.92	[1, 25]	3.08	8.19	1.40	0.00	0.04	10.26	2.32	4.35	5.48	1.60
9	J	8.15	[1,21]	10.47	2.22	0.02	3.09	0.04	5.96	7.64	0.54	4.25	11.51
10	M	6.91	[1,21]	2.34	5.43	0.07	0.00	0.09	18.61	1.39	3.28	0.66	0.77
11	N	8.41	[2, 22]	8.68	8.03	0.00	0.43	0.16	9.91	1.27	2.60	2.64	3.27
12	T	8.63	[0,24]	5.67	6.41	3.98	2.19	9.31	2.35	5.56	8.01	7.88	6.00

B.3. Detailed input of the patient waiting list

Table B.3 details the characteristics of the patient waiting list used as input in the case study. The table includes information about the patient ID, i, his/her type according to the diagnosis, s, the room of allocation, r, the benefit of admission, ζ_{is}, and penalty of overstay, θ_{is}.

TABLE B.3. Waiting list input data of the patient admission.

Patient, i	Type, s	Room, r	Benefit of admission, ζ_{is}	Penalty overstay, θ_{is}	Patient, i	Type, s	Room, r	Benefit of admission, ζ_{is}	Penalty overstay, θ_{is}
1	5	2	0.847	0.085	31	6	1	0.579	0.058
2	2	5	0.216	0.022	32	6	2	1.322	0.132
3	11	1	0.761	0.076	33	2	5	0.409	0.041
4	4	1	1.671	0.167	34	6	2	0.643	0.064
5	6	2	0.982	0.098	35	1	1	1.548	0.155
6	6	2	0.428	0.043	36	1	2	1.284	0.128
7	10	1	0.869	0.087	37	5	2	0.535	0.053
8	4	1	0.537	0.054	38	8	2	1.120	0.112
9	4	1	0.809	0.081	39	10	2	0.755	0.076
10	1	1	0.927	0.093	40	12	2	2.091	0.209
11	4	2	0.932	0.093	41	4	7	1.365	0.136
12	7	1	0.642	0.064	42	3	3	0.486	0.049
13	5	1	0.594	0.059	43	7	1	0.795	0.080
14	12	1	0.467	0.047	44	8	1	0.266	0.027
15	4	1	0.624	0.062	45	5	2	0.736	0.074
16	1	2	1.838	0.184	46	3	3	0.507	0.051
17	5	2	0.825	0.082	47	10	6	1.310	0.131
18	4	1	0.479	0.048	48	5	8	0.962	0.096
19	9	1	1.472	0.147	49	6	2	0.511	0.051
20	8	2	1.753	0.175	50	6	3	0.927	0.093
21	9	1	0.869	0.087	51	1	9	3.094	0.309
22	12	2	1.062	0.106	52	6	2	1.191	0.119
23	5	6	1.198	0.120	53	4	1	0.406	0.041
24	5	2	0.969	0.097	54	6	3	0.207	0.021
25	11	1	1.235	0.124	55	9	1	0.843	0.084
26	5	1	0.876	0.088	56	9	2	1.059	0.106
27	12	1	0.939	0.094	57	4	1	0.471	0.047
28	11	2	0.815	0.082	58	11	2	1.180	0.118
29	11	6	0.446	0.045	59	1	4	2.915	0.291
30	1	2	1.494	0.149	60	4	10	4.851	0.485

B.4. Maximum budget of overstay

Table B.4 describes the data employed to define the maximum value of overstay, O_r^{max}. Columns 1 and 2 display the care unit/room ID and description of each room in the hospital under study. Column 3 indicates the parameters assumed for each room.

TABLE B.4. Distribution parameters of the maximum value of overstay per room.

Room ID	Description	O_r^{max} (days)
1	Medicine	60
2	Surgery	70
3	Ophthalmology	30
4	Adult transplant	20
5	Psychiatry	20
6	Adult pensioner	20
7	ICU	20
8	CMA	20
9	Intermediate care, Surgery	20
10	Intermediate care, Medicine	20

B.5. Distribution parameters of the patient LoS for the out-of-sample analysis

Table B.5 describes the distribution fitting of the patient's LoS described in Subsection 5.4.5. We employed historical data and assumed the LoS of the patient's type is drawn from different distributions. A Monte Carlo simulation was performed to generate the samples. We considered the length of stay is independent and identically distributed for each patient type.

TABLE B.5. Length of Stay input parameters of the probability distribution mix for the out-of-sample test analysis.

Patient type	Sample size	MIX PROB 1	MIX PROB 2	MIX PROB 3	Exponential distribution
C	5215	Gamma [1.67, 5.94]	Weibull [10.68, 1.37]	Exponential [9.92]	Exponential [5.05]
F	1919	Weibull [14.76, 1.87]	Exponential [13.01]	Chi-square [13]	Exponential [13.01]
H	2407	Exponential [5.05];	Chi-square [4]	Gamma [0.97, 5.24]	Exponential [5.05]
I	8684	Weibull [12.66, 1.35]	Chi-square [11]	Lognormal [2.15,0.84]	Exponential [11.26]
K	9071	Chi-square [5]	Gamma [1.13, 5.90]	Lognormal [1.45,0.99]	Exponential [6.71]
S	8874	Lognormal [1.87,0.93]	Exponential [9.17]	Weibull [10.15, 1.27]	Exponential [9.17]
Z	1653	Gamma [1.88, 2.99]	Lognormal [1.56,0.62]	Exponential [5.25]	Exponential [5.25]
E	2966	Normal [7.22, 8.91]	Log-logistic [1.89, 6.43]	Burr [104.3,1.31,23.04]	Exponential [8.91]
J	3274	Lognormal [1.78,0.89]	Logistic [3.51, 8.15]	Normal [6.37, 8.15]	Exponential [8.84]
M	2523	Log-logistic [1.86, 5.144]	Normal [6.37, 6.91]	Gamma [1.17,5.87]	Exponential [6.91]
N	4377	Logistic [3.50, 8.41]	Burr [10.23,1.89,1.86]	Lognormal [1.85,0.77]	Exponential [8.41]
T	3243	Burr [4671.8,1.29,2767.3]	Gamma [1.355, 6.36]	Log-logistic [1.83,6.91]	Exponential [8.63]