A Robust Optimization Approach for Advance Scheduling in Health Care Systems with Demand Uncertainty: Policy Insights

by

Nafise Niazi Shahraki

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Sciences
in
Management Sciences.

Waterloo, Ontario, Canada, 2020

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

Supervisors: Houra Mahmoudzadeh
Assistant Professor, Department of Management Sciences,
Faculty of Engineering, University of Waterloo

Hossein Abouee Mehrizi
Associate Professor, Department of Management Sciences,
Faculty of Engineering, University of Waterloo

Other Members: James H. Bookbinder
Professor, Department of Management Sciences,
Faculty of Engineering, University of Waterloo

Sibel Alumur Alev
Associate Professor, Department of Management Sciences,
Faculty of Engineering, University of Waterloo

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Patient wait times have increased significantly over the past few decades. According to the Canadian Institute for Health Information (CIHI), 40% of Canadians have experienced difficulties in receiving diagnostics tests. MRI wait times have increased by 26% from 2012 to 2016. The lengthy wait times for the health care systems are translated to economic losses and risks to the lives of Canadians. These inefficiencies in health care systems are an indication that health care infrastructure investment has not been able to keep pace with the increased demands. While building new health care infrastructure to create capacity may be the first solution that comes to mind, it is often not feasible due to budget limitations. Optimizing the use of the existing capacity is a more feasible and cost-effective solution to healthcare system inefficiencies. This research builds on previous literature and proposes a robust optimization method for a multi-priority multi-period advance scheduling problem with wait time target which is solved using a proposed adversarial-based algorithm. A sensitivity analysis is conducted to calibrate the model parameters. Several numerical examples are used to extract practical policy insights. The advantages of the robust model in comparison with the deterministic model are highlighted. It is shown that the robust modelling leads to policies that are easier to execute and are more suitable for policy planning purposes when compared to deterministic modelling.

Acknowledgements

First of all, I would like to express my special thanks to my supervisors, Professor Houra Mahmoudzadeh and Professor Hossein Abouee Mehrizi for their continuous support, encouragement, and guidance during my graduate studies.

I would like to thank my committee members, Professor Sibel Alumur and Professor James Bookbinder for taking the time to read my thesis and giving me valuable feedback to improve this work.

I am also thankful to my mom and dad for their unconditional love and support during my life. They are the ones who encouraged me to strive for excellence. My appreciation extends to my dear sisters, Sara and Elaheh, and my dear brother, Majid for being always very supportive and encouraging.

Last but not least, I would like to thank my loving husband, Hamed, without his support and help, I could not have done this. Hamed, I thank you from the deep of my heart for all you have done for me. You could not be more supportive.

Dedication

To Hamed,

To my beloved Mom,

To the memory of my Dad.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Advance Patient Scheduling under Demand Uncertainty

According to the Fraser Institute [2019], patients' wait times in Canada have increased by 173% from 1993 to 2019. Based on a survey conducted in 2015, 40% of patients in Canada have experienced difficulties in receiving diagnostic tests. The average 90^{th} percentile wait time for an MRI examination in Canada was 186 days in 2016, which shows an increase of 26% compared to the year 2012 (147.2 days) (CIHI [2016]). It is evident that while the demand for diagnostic tests has increased, health care facilities and infrastructure have not been able to keep pace with the increased demand.

Meeting the wait-time target - the time during which a patient should receive service - could be as low as 40% for some operation (CIHI [2017]); it is an indication of the healthcare system and infrastructure inefficiency given that the defined benchmarks are already high (e.g., 182 days for knee replacement (CIHI [2019])). Long wait times can have serious consequences for patients by potentially allowing reversible medical conditions to become chronic or irreversible conditions.

Two of the major causes of long wait times in Canada's health care system are limited capacity and poor patient scheduling. Increasing capacity can be a viable and ideal solution to the healthcare system inefficiencies, but it requires substantial capital investment and thus is costly and often not a feasible and practical solution in the short term. Another alternative is optimizing patient scheduling, which makes use of the existing infrastructure and provided capacity and is a cost-effective solution with short- and long-term benefits.

With respect to modelling of uncertainty, studies that aim at optimizing patient scheduling are categorized to deterministic, stochastic and robust modelling; the studies are also categorized to allocation and advanced scheduling depending on their resolution of scheduling. Allocation scheduling is about the allocation of patients to appointments start time throughout a specific day while advance scheduling deals with patient allocations to different days over a scheduling horizon. Each of these modelling approaches has its pros and cons, which are elaborated in the Literature Review section of this thesis.

While this research is indebted to previous studies in the literature for making a substantial advancement in the patient scheduling area, it extends the literature boundaries by relaxing some of the simplifying assumptions in previous studies to make patient scheduling modelling more practical. This study proposes a robust optimization method, which adopts the notion of "budget of uncertainty" introduced by Bertsimas and Sim [2004] and guarantees an optimal scheduling scheme for a multi-class multi-period patient scheduling problem under demand uncertainties, where total capacity is a decision variable in the model. While ensuring optimality of the scheduling scheme, the optimization problem formulation also guarantees wait time targets for patients of different priorities. The method adds methodological novelty to advanced scheduling by considering patient classes and associated wait time targets for each class. Modelling parameters are calibrated to reproduce real-world measures. The importance of robust modelling of patient scheduling problem is highlighted through a comparative analysis of system-wide performance measures for robust and deterministic modelling. Extensive numerical analysis is used to extract practical policies and insights. The developed methodology provides a powerful decision-making tool to be used by health care professionals for optimal advanced patient scheduling by taking into account demand uncertainties with the goal of providing an acceptable level of service for all patients while prioritizing urgent cases. Since the modelling approach will be general, the same methodology can be applied to different advance scheduling applications.

This thesis is organized as follows: Chapter 2 presents a review of the literature on advanced patient scheduling and identifies gaps based on which the objectives of this research is set. Chapter 3 presents the robust formulation of the advanced patient scheduling problem and outlines the adversarial solution algorithm, which is used to solve the optimization problem. Chapter 4 presents the sensitivity analysis for the proposed model and the results and numerical analysis, followed by identifying trends in the results and drawing practical policy insights. The thesis concludes with highlights of findings and future research avenues in Chapter 5.

Chapter 2

Literature Review

2.1 Advance Patient Scheduling

The patient scheduling problem has been studied extensively over the past decades. Seminal studies in this area are traced back to the works of Bailey [1952] and Lindley [1952]. Applications of patient scheduling range from physician appointment scheduling to more critical scheduling applications such as surgery scheduling and scheduling of diagnostic tests, including MRI and Radiation Therapy Sauré et al. [2020]. There are five literature review papers that are found useful (Magerlein and Martin [1978], Przasnyski [1986], Cayirli and Veral [2003], Cardoen et al. [2010], and Ahmadi-Javid et al. [2017]). For further details, the readers are referred to these studies.

The patient scheduling studies in the literature are categorized to advance patient scheduling and patient allocation scheduling studies. The former deals with the allocation of patients to appointments start time throughout a specific day while the latter deals with patient allocations to different days over a scheduling horizon (Truong [2015], Sauré et al.

[2020], Gocgun and Ghate [2012]).

This research falls into the first stream of "advance scheduling". In advance scheduling, various factors such as service time uncertainty, demand uncertainty, and no-show and cancellation can be incorporated into the model at the cost of adding complexities. The incorporation of these factors also affects the selection of the modelling approach. The two common modelling approaches for incorporating uncertainties in the advance patient scheduling problem are stochastic programming and robust optimization Addis et al. [2014b]. In this chapter, first, a review of a number of selected studies that used stochastic programming in the advance scheduling problem is provided in Section 2.2. Then, a brief review of robust optimization, its application in advance scheduling, and a number of relevant studies in robust optimization and patient scheduling are presented in Section 2.3.

2.2 Stochastic Advance Patient Scheduling

Stochastic programming is a mathematical approach, which is extensively used for modeling problems with uncertain parameters. The approach was first introduced by Dantzig [1955]. In Stochastic programming, a probability distribution is considered for a problem's uncertain parameters. In this section advance patient scheduling studies that use stochastic optimization approach are presented. Sub-section 2.2.1 presents a review of some selected studies in which service time is considered to be uncertain, while 2.2.2 presents an overview of some studies with demand uncertainty.

2.2.1 Service Time Uncertainty

Gerchak et al. [1996] studied an advance scheduling problem for a general surgery operating room which is used for two types of surgeries – elective surgery and emergency surgery- and the capacity utilization for either of surgeries is uncertain. The problem is modelled as a stochastic dynamic programming model with the aim to schedule elective surgeries during the time horizon. An optimal policy regarding the optimal level of reserved capacity for emergency patients is extracted and the performance of the proposed policy is evaluated through numerical examples. The numerical examples confirm that the optimal level of reserved capacity for emergency patients is mostly affected by the number of deferred surgeries.

Denton and Gupta [2003] modelled a single priority multi-period surgical patients appointment scheduling problem as a two-stage stochastic linear program. The proposed formulation is unique in that it offers flexibility in finding optimal start times under various cost structures including piece-wise linear cost structure for overtime, waiting and idling costs. The formulation can easily be extended to model customer no-show in applications where the no-show consideration matters. In this thesis, a decomposition-based approach is used for solving the deterministic equivalent of the scheduling problem. General upper and lower bounds for the problem are developed and used by a variant of the L-shaped algorithm (Van Slyke and Wets [1969]) to calculate the optimal solution to the problem. The practical importance of the proposed model is illustrated through the numerical examples.

Klassen and Rohleder [2004] studied a multi-period advance scheduling problem with two classes of patients, regular and urgent. Service time is considered unknown in this study. Studies prior to this research focus on a single period scheduling problem, while this study considers a dynamic, multi-period environment, which is more realistic. The two main objectives of this study are finding the best scheduling rule for multi-period advance scheduling problem in presence of urgent patients and finding the best placement of reserved time slots for serving urgent demand to ensure on-time service. In this study, the performance of low variance clients at the beginning of the schedule (LVBEG) rule, which was introduced for single-priority scheduling problem by Klassen and Rohleder [1996], is evaluated for a multi-period scheduling problem. It is shown that LVBEG rule has the highest performance when compared to other rules such as first call, first appointment (FCFA).

Chakraborty et al. [2010] studied a sequential scheduling problem for a multi-period single-priority scheduling, which builds on the myopic scheduling problem studied by Muthuraman and Lawley [2008] by incorporating arbitrary service time distribution. The authors overcome the intrinsic computational challenges of general service time distribution by approximating the service time by a gamma distribution. Using a data-set of observed service times from New York Metropolitan Hospital, the gamma approximation of service time is proven accurate when compared to memory-less exponential approximation. Through simulation, it is concluded that the formulation proposed by the authors outperforms other formulations in the literature in terms of expected profit and total number of scheduled patients.

Saure et al. [2012] studied the radiation therapy appointment scheduling problem. They aim at finding scheduling policies for radiation therapy appointments to cost effectively decrease patients wait times. The scheduling problem is modelled as a discounted infinite-horizon Markov decision process (MDP), which is solved using an approximate dynamic programming approach. In this research, multiple priority types of patients are considered in the model and patients can receive treatment in multiple days with varying treatment session length. Considering multiple treatment with varying length is what differentiate this study from that of Patrick et al. [2008]. The performance of proposed method and extracted policies are evaluated through numerical experiments using data from British Columbia Cancer Agency (BCCA).

Gocgun and Ghate [2012] studied different classes of dynamic scheduling problems in advance scheduling and patient allocation scheduling with multiple resources; this study is an extension to Patrick et al. [2008] that proposed a model for advance patient scheduling with a single resource. The problems in this study are modelled as a weakly-coupled MDP. An approximate dynamic programming method that uses Lagrangian Relaxation and Column Generation approach is developed to provide an approximate optimal solution to the MDP problems. By using the approximate solution of the weakly coupled MDP problems, a set of scheduling decisions are extracted and the performances of the proposed decisions is tested through simulation. The simulation results confirm that applying these Lagrangian decisions in advance scheduling is more beneficial in terms of profits and costs than in allocation scheduling.

Sauré et al. [2020] studied a multi-period, multi-class advance patient scheduling problem with stochastic service time. This study considers stochastic service time in an advance scheduling problem, which was not addressed prior to this study as it adds complexity to the problem by incorporating the idle-time and overtime of medical resources. The problem is modelled as an infinite-horizon MDP model for deterministic and stochastic service times. Policy analysis results show that policies of advance scheduling under deterministic service time do not perform well in advance scheduling with stochastic service time.

2.2.2 Demand Uncertainty

Lamiri et al. [2008] studied the operating room planning problem with two classes of patients - elective and emergency patients. The demand for emergency cases is considered random and uncertain, requiring patients to receive service upon arrival, while elective patients are scheduled for future days. The problem is modelled as a stochastic model with random emergency demand - a solution methodology that combines Monte-Carlo simulation and mixed integer programming is proposed. The numerical experiments provided in the research highlight the importance of incorporating uncertainty in modelling scheduling problems. The importance of uncertainty consideration is further highlighted by comparing the results of the proposed stochastic model in this study with that of a deterministic model, where uncertainty is disregarded.

Patrick et al. [2008] studied a multi-priority patient dynamic scheduling for a diagnostic facility in a public health-care setting, where patients of different priorities had different wait-time targets. The problem is modelled as a discounted infinite-horizon MDP. The cost function in the MDP model is constituted of three different components: costs associated with waiting beyond the wait-time target, costs associated with overtime/surge capacity, and penalty costs associated with existing demand that never received service. Given the large size of the state space, finding a direct solution is not possible and thus, an approximation of the dynamic programming is used to solve the problem. The approximate solution of the problem results in an approximate optimal booking policy. The optimality of proposed booking policy is tested through simulation, which illustrates that the optimal booking policy is robust to changes in cost parameters in the cost function.

Liu et al. [2010] developed a model for single-priority, single-period, dynamic appointment scheduling problem with consideration of no-show and cancellation possibilities. The probabilities of no-show and cancellation are assumed to be increasing with delays. The model parameters are calibrated to simulate real-world data collected from a clinic. The study compares the performances of the developed scheduling policies (OTPSP, Imp-OAP, and Imp-OTPSP) against a few benchmark scheduling policies (open access policy, threshold policy, balanced scheduling policy, random scheduling policy) in terms of accuracy performance and concluded that the developed scheduling policies of this research outperforms the benchmark policies.

Gocgun and Puterman [2014] studied the problem of dynamic patient scheduling for a chemotherapy centre. Two types of appointments, follow-up (FU) and new patient (NP), are considered in this study and patients are considered to have different target dates and tolerance limits. Patient types are identified based on their tolerance limit. The problem is modelled as an MDP and an approximate dynamic programming method is used to find the solution of the MDP. Using the approximate dynamic programming solution to the problem, scheduling policies and booking rules are extracted; the performance of these extracted policies are compared to that of practical and easy-to-use heuristic booking rules and it is concluded that overall, their proposed policies outperform the easy-to-use heuristic booking rules.

Astaraky and Patrick [2015] proposed a dynamic programming model for a multi-class multi-resource surgical scheduling problem. The objective of this study is to provide a surgical scheduling policy by which not only the operating rooms and recovery beds are used efficiently, but also the lead time between the day of patients' request for surgery and the day of surgery, i.e., patients' wait time, is minimized. The problem is modelled as an MDP problem and solved by a developed algorithm based on Least Squares Approximate Policy. The applicability of the proposed formulation of the problem and the proposed solution method is validated using real-world data of a hospital.

Parizi and Ghate [2016] studied a multi-class, multi-resource advance patient scheduling with cancellation, now-shows, and overbooking. A weakly-coupled MDP formulation is developed for the scheduling problem. It is shown that the Lagrangian relaxation with constraint generation and affine value function approximation method introduced by Gocgun and Ghate [2012] can be applied to the weakly coupled MDP problems. Lagrangian policies are extracted in this study and are compared to myopic heuristics through numerical experiments; the analysis concludes that the proposed policies outperform the myopic heuristics.

Jiang et al. [2020] studied the MRI patient scheduling problem. A descriptive analysis is conducted on MRI scheduling data from 74 hospitals in order to extract system characteristic parameters and demand. To decrease the number of days that patients would have to wait beyond their wait-time target without increasing the capacity, two dynamic scheduling policies are proposed, modelled and evaluated: weight accumulation policy and priority promotion policy. A simulation model for multi-priority advance patient scheduling with wait-time targets is developed using the extracted modelling parameters and demand and the performance of the proposed policies are evaluated. comparing the performances of the two proposed policies with that of the current practice, it is concluded that the proposed policies can improve the performance of the system.

Klassen and Rohleder [1996] studied a dynamic outpatient appointment scheduling problem with different types of patients. They address the dynamic patient scheduling problem and compare different scheduling rules to minimize the wait time of the clients and the idle time of the service providers. In the formulation of the problem, both patient's service time and demand are considered uncertain. The study simulates a variety of rules under different conditions and concludes that the best patient scheduling rule is driven by the performance measure objectives of a clinic as well as the environmental factors. The study concludes that acceptable performances can be achieved when patients with larger service time deviations are scheduled toward the end of the scheduling horizon.

2.3 Robust Advance Patient Scheduling

Robust Optimization (RO) is a modelling methodology for managing and solving optimization problems with data uncertainty. RO was proposed in the early 1970s by Soyster [1973] and has been extensively studied since then. For comprehensive reviews of robust optimization, readers are referred to Beyer and Sendhoff [2007], Bertsimas et al. [2011], Ben-Tal et al. [2009], and Gabrel et al. [2014]. As apposed to stochastic programming that requires knowledge about uncertainty distribution of the uncertain parameters, RO only requires limited information for the uncertainty range of the uncertain parameters. In RO, uncertainty is often expressed by an uncertainty set that includes all possible scenarios of the uncertain data.

One of the concerns in RO is that the optimal solution obtained in a robust model might be too conservative. Soyster [1973] developed a linear optimization model to construct an optimal solution that remains feasible for all possible scenarios of the uncertain data, which leads to a too conservative optimal solution. El Ghaoui and Lebret [1997] and Ben-Tal and Nemirovski [2000] proposed less conservative robust modeling approaches by defining the ellipsoidal uncertainty sets. Then, Bertsimas and Sim [2004] proposed a new robust approach to address the issue of over conservatism. This approach offers full control over the level of uncertainty in the model by introducing the notion of “budget of uncertainty” which is a parameter in the model that determines the maximum amount of uncertainty that is allowed in the model.

In the following section, a brief review of a number of selected studies in the area of robust optimization application in patient scheduling and capacity allocation problem are provided, which is followed by identifying the gaps in the literature and the contribution of this thesis.

Holte and Mannino [2013] conducted research on a robust advance patient scheduling problem with demand uncertainty. The complexity of their proposed model stems from considering the cyclic nature of patient allocation and demand uncertainty in their model. An algorithm based on the row- and column-generation developed and proposed as a solution methodology in this study. The convergence of the proposed algorithm is shown through simulation. The proposed model and solution algorithm are tested through real-world data collected from a major hospital in the city of Oslo in order to compute the master surgery schedules. It is concluded that under accurate estimation of demand, their proposed approach can provide reliable and efficient master surgery scheduling schemes.

Addis et al. [2014a] studied the operating room planning problem with uncertain surgery duration. Five classes of patients are considered in the model and each patient is assumed to have a different wait-time target. Limited overtime capacity is permitted in their proposed model and robust optimization is used to model the problem. The objective function is to minimize the penalty cost associated with the wait time and the level of urgency and tardiness of patients. In order to control the level of uncertainty in surgery duration in the problem, the notion of the budget of uncertainty introduced by Bertsimas and Sim [2004] has been used. Four different formulation versions of this problem are discussed and compared in this study, including the deterministic model without overtime possibility, the robust model without overtime possibility, the deterministic model with overtime possibility, and the robust model with overtime possibility. The quality of the obtained solution of the proposed formulations is evaluated through numerical examples whose parameters are calibrated based on real-life data. Then, the optimal patient assignments obtained from optimization models are tested through a set of randomly generated scenarios and results are compared. Results confirm that the robust models outperform the deterministic ones.

Aslani et al. [2020] studied the robust optimization model for tactical capacity planning in an outpatient setting. The two types of appointments considered in this study are: First visit (FV) and re-visit (RV). Each class of patients has a different access target (wait-time target) and demand is considered to be uncertain. The main source of demand uncertainty in this study is from FV patients. The objective function is to find the minimum required physician capacity such that each patient class receives service within their access targets. They apply the notion of the budget of uncertainty to control the level of demand uncertainty in the model. The problem formulation follows the robust approach introduced by Bertsimas and Sim [2004]. In their proposed model, patients can receive service with delay; the maximum number of postponed patients is a parameter, which is calculated using historical data. In this study, no assumption regarding the unmet demand is made in the proposed model.

Mahmoudzadeh et al. [2020] studied a robust multi-period multi-class, multi-priority advance patients scheduling with demand uncertainty. In their proposed model, each group of patients has a different wait-time target and required service time. For a given capacity, patients of different priorities are scheduled across the time horizon such that a certain level of service is achieved for each priority level. The level of conservatism in their proposed model is controlled by the notion of the budget of uncertainty. The proposed robust model is solved by using a two-stage reformulation approach. The proposed robust-based is compared to an equivalent deterministic model and it is concluded the robust model provides higher service levels and less wait times.

2.4 Contribution

In this thesis, the robust optimization approach is applied to a multi-class multi-period multi-priority advance patient scheduling with demand uncertainty, where over-time capacity is available by extra cost. The model proposed in this research does not consider service time uncertainty due to the nature of the MRI examination, i.e., fixed examination length for each level of priority is considered in the model. To the best of my knowledge, this problem has not been addressed by previous research in the literature. Among all of the studies which are reviewed in the reviewed literature, Aslani et al. [2020] and Mahmoudzadeh et al. [2020] are the most relevant ones to this work and the differences between the two studies and this research are explained below:

- Total capacity is a decision variable in the proposed model to make the model more realistic in representing real-world scenarios. A scenario that cannot be modelled without this consideration is hospitals with the option of overtime working hours to manage wait times.

- Demand uncertainty is considered for all priority levels. The level of uncertainty can be controlled by the notion of the budget of uncertainty, which is estimated by real-world data.

- A multi-objective function is defined in my proposed model. This multi-objective function enables the model to perform a trade-off analysis among different cost components. It also gives the decision maker the flexibility to choose cost function parameters to meet a system's operational goals.

- In the proposed model, due to the presence of uncertainty across rows and columns in the model, the solution methodology introduced by Bertsimas and Sim [2004] (for row-wise) is not applicable. Here, an adversarial algorithm is applied to solve the problem and handle the row-wise and column-wise uncertainty in the model.

Chapter 3

Methodology and Modeling

This chapter presents the mathematical tools used to model the multi-priority multi-period advance patient scheduling problem with demand uncertainty. The chapter is broken down as follows: Section 3.1 provides the problem description. Section 3.2 includes the mathematical formulation of the problem and explains the different cost components included in the objective function. In Section 3.3, an algorithm is developed for solving the problem and its convergence is discussed.

3.1 Problem Definition

In advance patient scheduling problem, a mismatch between the available capacity and patient demand results in long wait times. The complexities in advance patient scheduling problem is due to:

- Limited capacity and the costly nature of overtime capacity

- Demand uncertainty
- Service time uncertainty

This research considers a robust multi-period, multi-priority advance patient scheduling problem with demand uncertainty. In the proposed framework, patients are classified into multiple priority levels with different wait-time targets. Demand uncertainty is incorporated in the problem and is modelled using the notion of the budget of uncertainty introduced by Bertsimas and Sim [2004]. Total capacity in each period is a decision variable, but the regular capacity in each period is given. The objective is to minimize the overall cost, including the cost associated with capacity, the penalty cost associated with waiting beyond the wait-time target for each priority level, and the penalty cost associated with unmet demand, i.e., the patients who never receive service within the time horizon due to insufficient capacity. Total capacity is considered as a decision variable in the problem since some hospitals may work extra hours to manage their wait time. More details about the proposed framework for the studied problem is provided in the following sections.

3.2 Mathematical Formulation

In this section, a mathematical formulation for the problem is developed. Section 3.2.1 introduces the notations used in the mathematical formulations presented in this study. Section 3.2.2 provides information about how the demand uncertainty is incorporated in the model. Then, additional information about the structure of the objective function is presented in Section 3.2.3. Section 3.2.4 and 3.2.5 present the mathematical formulations of the deterministic and the robust models proposed in this study.

3.2.1 Notation

Sets

- $\mathcal{T} = \{1, \ldots, T\}$: Planning horizon
- $\mathcal{P} = \{1, \ldots, P\}$: Priority levels of patients
- $\mathcal{K} = \{1, \ldots, K\}$: Index set of linear segments in PWLC capacity cost function

Parameters

- $T =$ Last period in the planning horizon
- $P =$ Number of priority levels
- $\bar{d}_{t,p}$: Nominal value for demand of patients with priority level p at day t, $\quad p \in \mathcal{P}, t \in \mathcal{T}$
- $\hat{d}_{t,p}$: The maximum possible deviation of demand of patients with priority level p at day t, $\quad p \in \mathcal{P}, t \in \mathcal{T}$
- $\tilde{d}_{t,p}$: Total uncertain demand of patients with priority level p at day t, $\quad p \in \mathcal{P}, t \in \mathcal{T}$
- K : Number of segments in PWLC capacity cost function
- CL_i : i^{th} Capacity limit, $\quad i \in \mathcal{K}$
- θ_i : Capacity cost associated with i^{th} segment in the PWLC capacity cost function, $i \in \mathcal{K}$

- s_p : Service time of patients with priority level p, $\quad p \in \mathcal{P}$

- w_p : Wait-time target of patients with priority level p, $\quad p \in \mathcal{P}$

- Γ_p : Budget of uncertainty for patients with priority level p, $\quad p \in \mathcal{P}$

- $g_{t,p}$: Penalty cost for patients with priority level p who arrive at day t and do not receive service within the time horizon, $\quad p \in \mathcal{P}, t \in \mathcal{T}$

- $f_{t,n,p}$: Penalty cost of patients with priority level p who receive service at day t after waiting for n days, $\quad p \in \mathcal{P},\ t \in \mathcal{T},\ n \in \{1, \ldots, t\}$

Decision variable

- $x_{t,n,p}$: Integer decision variable denoting the number of patients with priority level p who receive service at day t after waiting for n days.

3.2.2 Demand Uncertainty

In the proposed robust model, demand is considered to be uncertain. In order to control the level of robustness in the model, demand uncertainty is incorporated to the model by using the notion of the budget of uncertainty introduced by Bertsimas and Sim [2004]. The parameter $\tilde{d}_{t,p}$ is the demand of patients with priority p who arrive at day t. Parameter $\tilde{d}_{t,p}$ is uncertain and belongs to an interval whose center is the nominal demand, $\bar{d}_{t,p}$, with the maximum variation of $\hat{d}_{t,p}$. Demand uncertainty can be controlled by the budget of uncertainty, Γ_p. The parameter Γ_p controls the level of uncertainty in the problem and it can take values between 0 and the length of the time horizon, T. In robust optimization, there is a trade-off between the level of robustness and the quality of the optimal solution.

Using larger values of Γ_p in the model leads to a higher level of conservatism in the model. It is necessary to use experts or decision-makers' opinion to determine the appropriate Γ_p in the model.

3.2.3 Structure of the Objective Function

In this section, the different cost components in the objective function are explained. These cost components can be categorized into three different types: (i) Penalty cost associated with unmet demand (C_U) (ii) Capacity cost (C_C) (iii) Penalty cost associated with patients who have been waiting beyond their wait-time target (C_D). Total cost (C_T) is the sum of the three cost components, C_C, C_D, and C_U which are explained in the subsequent sections.

(i) Capacity cost (C_C)

The level of capacity on the day t refers to the maximum number of appointment slots available at day t. An insufficient level of capacity is one of the factors that contribute to patients' long wait times. Although increasing the capacity would decrease patients' wait times, it is not always possible to increase capacity due to budget constraints. A distinguishing feature of the proposed model is that total capacity is a decision variable that enables the model to make a trade-off between the capacity cost incurred by increasing the capacity and penalty cost incurred by unmet demand and delayed patients.

A classic/common approach for defining the capacity cost function is to consider a linear function. The key problem with the linear capacity cost function (LCCF) is that the optimal solution under an LCCF is evident even without solving the model: the model would either increase the capacity to be able to serve all the patients at their arrival day or keep the capacity level at zero and do not serve any of patients at all. Therefore, a piece-wise linear convex (PWLC) function is proposed to capture the cost of the extra capacity.

In practice, in health service centers, there is usually a default level of capacity available to use. This level of capacity is to represent the available capacity under normal operating hours without requiring any additional resources. I assumed that the cost associated with the capacity level at this interval would be θ_1. However, there is also a limited over-time capacity available when the demand increases drastically. Each unit of over-time capacity causes extra cost (θ_2) in the system. This extra cost stems from additional working hours of staff and machines, or borrowing spaces from other departments. In some cases, increasing the capacity , even more, is also possible by additional cost (θ_3). This may include costs related to hiring new staff, constructing new spaces, and purchasing new machines. Therefore, it makes practical sense to assume that $\theta_1 < \theta_2 < \theta_3$. Figure 3.1 shows how capacity cost function is defined. Based on Figure 3.1, if the level of capacity falls within the range of $[0, CL_1]$, the capacity cost for each unit of capacity is θ_1 (which is considered equal to zero in Figure 3.1). If the level of capacity is in the range of $[CL_1, CL_2]$, for each unit of capacity beyond CL_1, a cost of θ_2 is considered in the model. Similarly, if the level of capacity falls within $[CL_2, CL_3]$, for each unit of capacity beyond CL_2, an extra cost of θ_3 is considered.

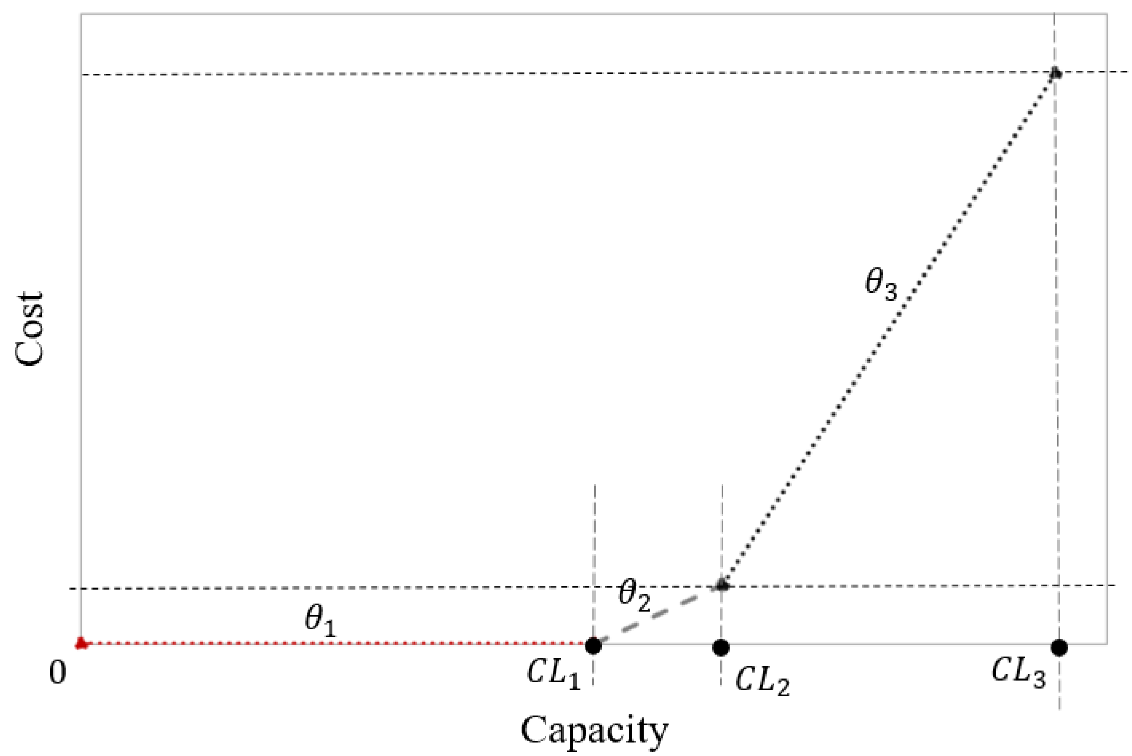


Figure 3.1: Capacity Cost Function

(ii) Penalty Cost associated with unmet demand (C_U)

As discussed in the previous section, the problem is defined in a way that capacity can be increased to provide service to patients of different priorities by incurring some cost. The other cost components considered in the objective function are the penalty cost for unmet demand and late service. By considering the penalty cost for unmet demand which refers to the penalty cost of not serving patients, the model is enforced to fix the level of capacity and unmet demand based on the trade-off between their corresponding costs. Thus, there is a balance between the capacity cost and delayed/unmet demand cost.

let $U_{t,p}$ denote the number of patients with priority p who arrive at day t and never receive service until the end of horizon. Total demand of patients with priority p at day t is equal to $\tilde{d}_{t,p}$. The value of $U_{t,p}$ is calculated by subtracting the total number of patients with priority p who arrive at day t and are assigned to different days during the time horizon for receiving service from the total demand. Equation 3.1 demonstrates how ($U_{t,p}$) is calculated.

$$U_{t,p} = \big(\tilde{d}_{t,p} \ - (\sum_{i=0}^{T-t} x_{T-i,T-t-i+1,p})\big) \tag{3.1}$$

There are different approaches for defining the penalty cost for unmet demand for each priority. One common approach is to define it as a function of the length of horizon (T), patient's priority level (p), arrival (t), and its wait-time target (w_p). I propose the following penalty cost:

$$g_{t,p} = Max\ (0,\ \frac{T-t-w_p+2}{w_p}) \tag{3.2}$$

Hence, the unmet demand penalty cost for different arrivals and different priorities can be written as follows:

$$\sum_{p=1}^{P}\sum_{t=1}^{T} U_{t,p} g_{t,p} = \sum_{p=1}^{P}\sum_{t=1}^{T} \big(\tilde{d}_{t,p} \ - (\sum_{i=0}^{n-1} x_{T-i,n-i,p})\big) \times \frac{T-t-w_p+2}{w_p} \tag{3.3}$$

(iii) Penalty cost for waiting beyond the wait-time target (C_D)

The third cost component in the objective function is the penalty cost for waiting beyond the wait-time target. For each priority level, the wait-time target (w_p) is defined which shows that patients of priority level p should receive service within their specific target.

For example, if the wait-time target of patients with priority 2 is 2 days, then the allowable delay for providing service to these patients is 2 days and if they do not receive service within their wait-time target, it creates a penalty cost in the model. For determining a wait-time target for each priority level, it is assumed that patients with higher acuity levels have a smaller wait-time target. If patients receive service within their wait-time target, then the penalty cost coefficient would be zero. For patients who are waiting longer than their wait-time target, w_p, a penalty cost will be considered. Equation (3.4) shows how the penalty cost coefficient is calculated for patients who arrive at day t and receive service at day y after waiting for $n = y - t + 1$ days, while their wait-time target is w_p.

$$f_{y,n,p} = \begin{cases} 0, & y \leq (t + w_p - 1) \\ y - t + 1 - w_p, & y \geq t + w_p \end{cases} \tag{3.4}$$

Table 3.1 shows an example for $f_{y,n,p}$ for different priority levels, different arrivals, and service days. In this example, patients are classified into 4 priority levels and the wait-time target for patients of priority 1, 2, 3, and 4 are 1, 2, 5, and 10 days, respectively. In Table 3.1, $f_{y,n,p} = 0$ for patients who receive service within their wait-time targets. For the same arrival and the same service day, $f_{y,n,p}$ is also increasing by the priority level which stems from the fact that higher priority levels have smaller w_p.

Table 3.1: An example of $f_{y,n,p}$ for different priority levels, arrivals, service days

Priority level	w_p	Arrival(t)	$f_{y,n,p}$				
			Service Day (y)				
			1	2	5	6	10
P1	1	Day 1	0	1	4	5	9
		Day 5	-	-	0	1	5
P2	2	Day 1	0	0	3	4	8
		Day 5	-	-	0	0	4
P3	5	Day 1	0	0	0	1	5
		Day 5	-	-	0	0	1
P4	10	Day 1	0	0	0	0	0
		Day 5	-	-	0	0	0

3.2.4 Deterministic Model

Two sets of non-negative integer decision variables in the model are considered, including:

- $x_{t,n,p}$ denoting the number of patients with priority level p who receive service at day t after waiting for n days, $\quad t \in \mathcal{T}, n \in \{1, \ldots, t\}, p \in \mathcal{P}$
- $c_{i,t}$ represents the amount of capacity at day t which is available with the cost equal to the cost associated with i^{th} segment in the PWLC capacity cost function, $\quad i \in \mathcal{K}, t \in \mathcal{T}$

A set of binary decision variables (indicator decision variable), considered in the model:

- $q_{i,t}$ which is an indicator decision variable for determining the range of the capacity level:

$$q_{i,t} = \begin{cases} 1, & \text{if } CL_i \leq \sum_{i=1}^{k} c_{i,t} \\ 0, & \text{otherwise} \end{cases} \quad , \forall i \in \mathcal{K},\ t \in \mathcal{T}$$

The proposed deterministic mathematical formulations is:

$$\begin{aligned} &\underset{x,c}{Min} \quad \sum_{p=1}^{P}\sum_{t=1}^{T}\big(\bar{d}_{t,p} \ - (\sum_{i=0}^{T-t} x_{T-i,T-t-i+1,p})\big) g_{t,p} \\ &+ \sum_{t=1}^{T}\sum_{i=1}^{k}\theta_i c_{i,t} \\ &+ \sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{n=1}^{t} x_{t,n,p} f_{t,n,p} \\ &s.t. \end{aligned} \tag{3.5}$$

$$\sum_{i=0}^{n-1} x_{t-i,n-i,p} \leqslant \bar{d}_{t-n+1,p} \qquad \forall t \in \mathcal{T}, \forall n \in (1,\ldots,t), \forall p \in \mathcal{P} \tag{3.6}$$

$$\sum_{p=1}^{P}\sum_{n=1}^{t} s_p x_{t,n,p} \leqslant \sum_{i=1}^{k} c_{i,t} \qquad \forall t \in \mathcal{T} \tag{3.7}$$

$$CL_1 q_{1,t} \leqslant c_{1,t} \leqslant CL_1 \qquad \forall t \in \mathcal{T} \tag{3.8}$$

$$(CL_i - CL_{i-1}) q_{i,t} \leqslant c_{i,t} \leqslant (CL_i - CL_{i-1}) q_{i-1,t} \qquad \forall t \in \mathcal{T}, i \in \{2,\ldots,k-1\} \tag{3.9}$$

$$0 \leqslant c_{k,t} \leqslant CL_k \qquad \forall t \in \mathcal{T} \tag{3.10}$$

$$\begin{aligned} &x_{t,p,n} \geq 0,\ integer \quad \forall t \in \mathcal{T},\ p \in \mathcal{P},\ n \in \{1,\ldots,t\} \\ &c_{i,t} \geq 0, \quad \forall i \in \mathcal{K}, \forall t \in \mathcal{T} \\ &q_{i,t} \in \{0,1\}, \quad \forall t \in \mathcal{T}, \forall i \in \mathcal{K} \end{aligned}$$

In the deterministic model, demand during the time horizon is known and is equal to nominal demand, $\bar{d}_t, \forall\, t \in \mathcal{T}$. The objective function (3.5) minimizes the total cost which is constituted of the penalty cost for unmet demand (C_U), capacity cost (C_C), and the penalty cost for patients who waited beyond their wait-time targets (C_D) as explained in Section 3.2.3. Constraint (3.6), ensures that total number of patients who arrive at day t and receive service during the time horizon does not exceed the demand at day t. Constraint (3.7) ensures that the total number of patients who receive service at day t, does not go beyond the capacity level at day t. Constraints (3.8), (3.9), and (3.10) together determine the capacity range at day t, $\forall t \in \mathcal{T}$.

3.2.5 Robust Model

Considering a fixed daily demand during the time horizon makes the model unrealistic. In order to incorporate demand uncertainty in the model, a robust approach is proposed. The decision variables in the robust model are as follows:

Non-negative decision variables:

- $x_{t,n,p}$ Integer decision variable denoting the number of patients with priority level p who receive service at day t after waiting for n days.[1]
- $c_{i,t}$ Integer decision variable representing the amount of capacity at day t which is available with the cost equal to the cost associated with i^{th} segment in the PWLC capacity cost function.
- $z_{t,p}$ Continuous decision variable, where $0 \leq z_{t,p} \leq 1$, denoting how much the demand of patients with priority p at day t is deviated from the nominal demand.

[1] $n \in \{1, 2, 3, ..., T\}$

Binary variable, $q_{i,t}$ which is an indicator decision variable for determining the range of the capacity level:

$$q_{i,t} = \begin{cases} 1, & \text{if } CL_i \leq \sum_{i=1}^{k} c_{i,t} \\ 0, & \text{otherwise} \end{cases} \quad , \forall i \in \mathcal{K},\ t \in \mathcal{T}$$

$$
\begin{aligned}
&\underset{x,c}{Min} \quad \underset{z}{Max} \quad \sum_{p=1}^{P}\sum_{t=1}^{T}\Big(\hat{d}_{t,p}z_{t,p} + \bar{d}_{t,p} \ - (\sum_{i=0}^{T-t} x_{T-i,T-t-i+1,p})\Big)g_{t,p} \\
&+\sum_{t=1}^{t=T}\sum_{i=1}^{k}\theta_i c_{i,t} + \sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{n=1}^{t} x_{t,n,p} f_{t,n,p} \qquad (3.11) \\
&s.t.
\end{aligned}
$$

$$\sum_{i=0}^{n-1} x_{t-i,n-i,p} \leqslant \bar{d}_{t-n+1,p} + z_{t-n+1,p}\hat{d}_{t-n+1,p} \qquad \forall t \in \mathcal{T}, \forall n \in (1,\ldots,t), \forall p \in \mathcal{P} \qquad (3.12)$$

$$\sum_{p=1}^{P}\sum_{n=1}^{t} s_p x_{t,n,p} \leqslant \sum_{i=1}^{k} c_{i,t} \qquad \forall t \in \mathcal{T} \qquad (3.13)$$

$$(CL_i - CL_{i-1})q_{i,t} \leqslant c_{i,t} \leqslant (CL_i - CL_{i-1})q_{i-1,t} \qquad \forall t \in \mathcal{T}, i \in \{2,\ldots,K-1\} \qquad (3.14)$$

$$CL_1 q_{1,t} \leqslant c_{1,t} \leqslant CL_1 \qquad \forall t \in \mathcal{T} \qquad (3.15)$$

$$0 \leqslant c_{k,t} \leqslant CL_k \qquad \forall t \in \mathcal{T} \qquad (3.16)$$

$$\sum_{t=1}^{T} z_{t,p} \leq \Gamma_p \qquad \forall p \in \mathcal{P} \qquad (3.17)$$

$$x_{t,p,n} \geq 0, \ integer \quad \forall t \in \mathcal{T}, \ p \in \mathcal{P}, \ n \in \{1,\ldots,t\}$$

$$c_{i,t} \geq 0, \quad \forall i \in \mathcal{K}, \forall t \in \mathcal{T}$$

$$0 \leq z_{t,p} \leq 1 \quad \forall t \in \mathcal{T}, \ p \in \mathcal{P}$$

$$q_{i,t} \in \{0,1\}, \quad \forall t \in \mathcal{T}, \forall i \in \mathcal{K}$$

In Section 3.2.2, the explanation regarding incorporating demand uncertainty in the robust model is provided. Considering demand uncertainty in the model affects the objective function of the deterministic model given in (3.5) and the Constraint (3.6). In the deterministic model, demand values are assumed to be known and certain. Thus, the objective function is to minimize the total cost, C_T, given that x and c are decision variables. However, in the robust model, the objective function is to minimize the total cost constituted of C_C, C_D, and C_U for the worst-case realization of demand within the budget of uncertainty, Γ_p, $\forall p \in P$. Thus, the objective function (3.11) is defined as a min-max function in which the total cost is minimized for the worst-case demand within the budget of uncertainty. Decision variable z determines on which days during the time horizon, demand has a deviation from the nominal demand.

Constraints (3.13), (3.14), (3.15), and (3.16) in the robust model are exactly the same as Constraints (3.7), (3.8), (3.9), and (3.10) of the deterministic model, respectively. The role of Constraint (3.17) is to ensure that for each priority level p, the demand of at most Γ_p days during the time horizon will deviate from its nominal value, $\bar{d}_{t,p}$.

3.3 Solution Methodology

3.3.1 Introduction and Motivation

The robust model proposed in this study has two different types of decision variables: (i) the integer decision variables $x_{t,n,p}$ and $c_{i,t}$, (ii) the continuous decision variable $z_{p,t}$ which are explained in Section 3.2.5. Decision variable $z_{p,t}$ provides the model with the flexibility of spreading the uncertainty over the time horizon.

Looking at the objective function given in (3.11), right-hand-side of the Constraint (3.12), and Constraint (3.17), it can be noticed that the robust model has a combination of row-wise and column-wise uncertainty. Bertsimas and Sim [2004] propose a solution methodology for solving the robust problem in which the uncertainty is across rows and uncertain parameters in each row are independent of each other. Here, in our proposed model the uncertain parameter across rows and columns are not independent. Thus, the classic approach proposed by Bertsimas and Sim [2004] cannot be applied to solve the robust problem.

For solving this problem, an adversarial-based algorithm is proposed. In the following sections, more details about the solution methodology are provided.

3.3.2 An Adversarial-based Algorithm

In the proposed adversarial algorithm, first, the original problem is decomposed into two sub-problems: the adversarial sub-problem and the implementer sub-problem. Then, the optimal solution is obtained by solving the sub-problems iteratively, until they converge. The sub-problems are explained in the Section 3.3.3 and 3.3.4. The idea behind this algorithm is explained in Bienstock and ÖZbay [2008]. As mentioned before, the original robust optimization problem has three sets of decision variables: $x_{t,n,p}$, $c_{i,t}$, $z_{t,p}$. The implementer sub-problem is a simplified form of the original robust problem with two sets of decision variables: x and c. The optimal value for $x_{t,n,p}$ and $c_{i,t}$ in the implementer sub-problem are calculated for a given set of $z \in \mathcal{Z}$. At the first iteration of the algorithm $\mathcal{Z}$ has only a member, z_0. Since the implementer sub-problem is a relaxed version of the original problem, its optimal solution provides a lower bound for the original problem and the lower bound is always ascending. Thus, the lower bound, L, can be updated at each iteration. On the other hand, the adversarial sub-problem introduced in 3.3.4 is a maximization problem whose only decision variable is z. Using the optimal x^* and c^* obtained from the implementer sub-problem, a new optimal z, named z^*, is calculated from the adversarial sub-problem. At each iteration, an optimal z obtained from the adversarial sub-problem is added to $\mathcal{Z}$ and the upper bound, U, is also updated if the $f(x^*, c^*, z^*)$ is less than the current upper bound. Figure 3.2 shows how the adversarial algorithm works.

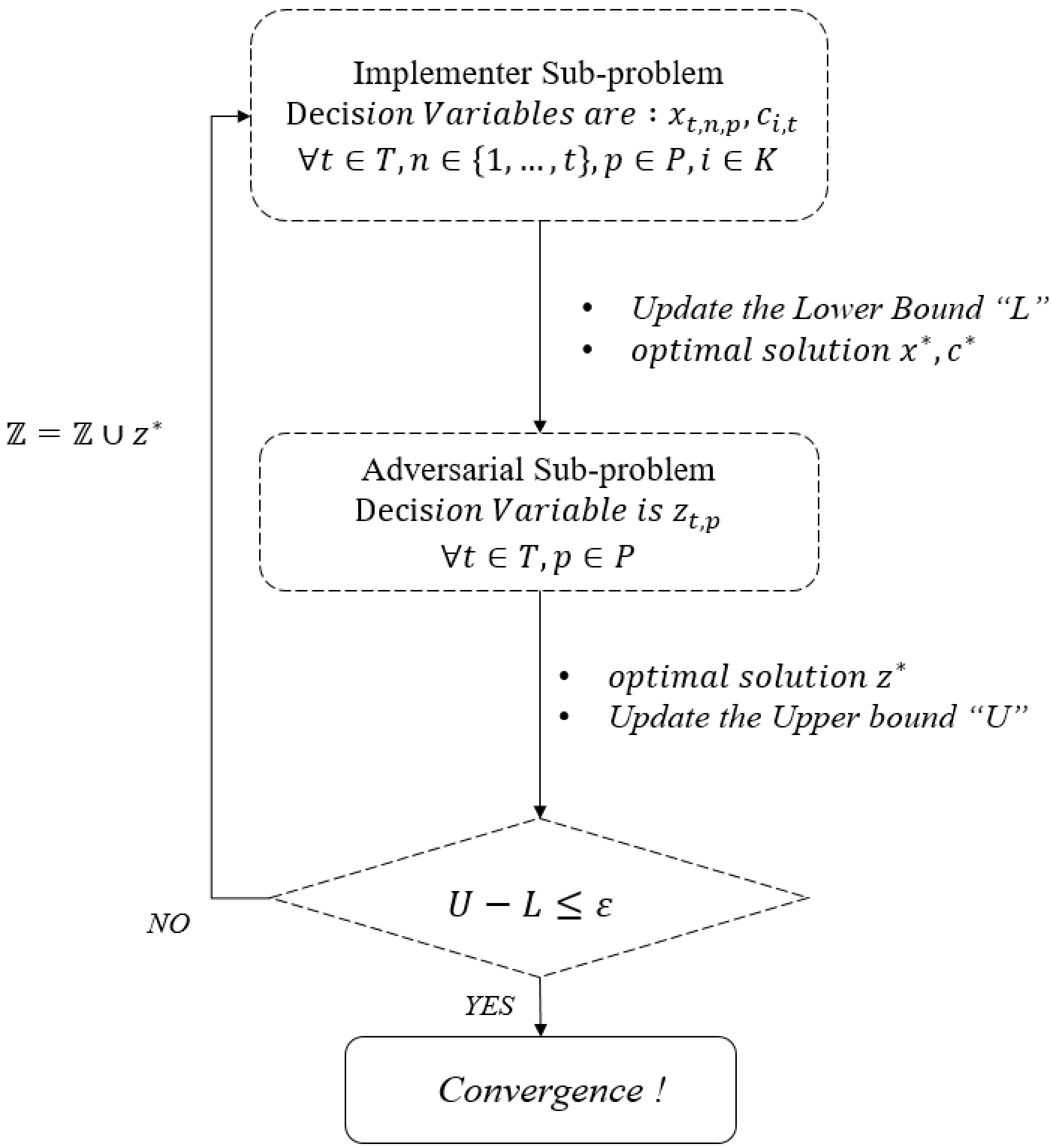


Figure 3.2: The Adversarial Algorithm Flow-chart

3.3.3 The Implementer Sub-problem

The implementer sub-problem is defined as follows:

$$\underset{\text{x,c}}{\text{Min}} \quad \underset{\text{z}\in\mathcal{Z}}{\text{Max}} \quad \sum_{p=1}^{P}\sum_{t=1}^{T}\left(\hat{d}_{t,p}z_{t,p}+\bar{d}_{t,p}-(\sum_{i=0}^{T-t}x_{T-i,T-t-i+1,p})\right)\times\frac{(T-i-w_p+2)}{w_p}$$

$$+\sum_{t=1}^{T}\sum_{i=1}^{k}\theta_i c_{i,t} \tag{3.18}$$

$$+\sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{n=1}^{t}x_{t,n,p}f_{t,n,p}$$

$s.t.$

$$\sum_{i=0}^{n-1}x_{t-i,n-i,p}\leqslant\bar{d}_{t-n+1,p}+z^*_{t-n,p}\hat{d}_{t-n,p} \qquad \forall p\in\mathcal{P},t\in\mathcal{T},\forall n\in\{1,\ldots,t\} \tag{3.19}$$

$$\sum_{p=1}^{P}\sum_{n=1}^{t}s_p x_{t,n,p}\leqslant\sum_{i=1}^{k}c_{i,t} \qquad \forall t\in\mathcal{T} \tag{3.20}$$

$$(CL_i-CL_{i-1})q_{i,t}\leqslant c_{i,t}\leqslant(CL_i-CL_{i-1})q_{i-1,t} \qquad \forall t\in\mathcal{T},i\in(2,k-1) \tag{3.21}$$

$$CL_1q_{1,t}\leqslant c_{1,t}\leqslant CL_1 \qquad \forall t\in\mathcal{T} \tag{3.22}$$

$$0\leqslant c_{k,t}\leqslant CL_k \qquad \forall t\in\mathcal{T} \tag{3.23}$$

$$x_{t,p,n}\geq 0,\ Integer \quad \forall t\in\mathcal{T},\ p\in\mathcal{P},\ n\in\{1,\ldots,t\}$$

$$c_{i,t}\geq 0, \quad \forall i\in\mathcal{K},\forall t\in\mathcal{T}$$

$$q_{i,t}\in\{0,1\}, \quad \forall t\in\mathcal{T},\ \forall i\in\mathcal{K}$$

The Implementer sub-problem is an integer programming problem in which $x_{t,n,p}$, $c_{i,t}$, and $q_{i,t}$ are decision variables and $z_{t,p}$ is assumed to be known and belongs to a known set

$\mathcal{Z}$. The set $\mathcal{Z}$ becomes updated at each iteration. The min-max objective function in the implementer sub-problem can be reformulated as:

$$\underset{x,c,A}{Min} \qquad A \tag{3.24}$$

$$
\begin{aligned}
& s.t. \\
A \geq \quad & \sum_{p=1}^{P}\sum_{t=1}^{T}\big(\hat{d}_{t,p}z_{t,p} + \bar{d}_{t,p} - (\sum_{i=0}^{T-t} x_{T-i,T-t-i+1,p})\big) \times \frac{(T-i-w_p+2)}{w_p} \\
& + \sum_{t=1}^{T}\sum_{i=1}^{k}\theta_i c_{i,t} \qquad\qquad\qquad (3.25)\\
& + \sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{n=1}^{t} x_{t,n,p} f_{t,n,p} \qquad\qquad \forall t \in \mathcal{T}, \forall p \in \mathcal{P}, \forall z \in \mathcal{Z}
\end{aligned}
$$

The Constraint (3.25), is to ensure that the objective function is minimized for the worst possible demand within the budget of uncertainty. Each new z obtained from solving the adversarial sub-problem in each iteration is added to the set $\mathcal{Z}$. It generates a cut in the formulation of the implementer sub-problem in the next iteration through the Constraint (3.25). It is similar to the idea of adding cuts at each iteration in Benders' Decomposition Algorithm. Constraint (3.19) ensures that the total number of patients of priority p who arrive at day t and receive service during the time horizon does not go beyond the demand of patients of that priority at day t. In this constraint, the value of z^* from the last iteration is used. Constraints (3.20) to (3.23) are exactly the same as Constraints (3.7) to (3.10) in the deterministic model which are included in the robust model as well.

3.3.4 The Adversarial Sub-problem

The adversarial sub-problem is an LP in which $z_{t,p}$ is a decision variable and $x_{t,n,p}$, $c_{i,t}$, $q_{i,t}$ are assumed to be known from solving the implementer sup-problem in the previous step.

The formulation of the adversarial sub-problem is as follows:

$$\underset{z}{Max} \quad \sum_{p=1}^{P}\sum_{t=1}^{T}\Big(\hat{d}_{t,p}z_{t,p} + \bar{d}_{t,p} \; - (\sum_{i=0}^{T-t} x_{T-i,T-t-i+1,p})\Big) \times \frac{(T-i-w_p+2)}{w_p}$$

$$+ \sum_{t=1}^{T}\sum_{i=1}^{k} c_{i,t} \tag{3.26}$$

$$+ \sum_{p=1}^{P}\sum_{t=1}^{T}\sum_{n=1}^{t} x_{t,n,p} f_{t,n,p}$$

$$s.t.$$

$$\sum_{i=0}^{n-1} x_{t-i,n-i,p} \leqslant \bar{d}_{t-n+1,p} + z_{t-n+1,p}\hat{d}_{t-n+1,p} \qquad \forall t \in \mathcal{T}, \forall n \in (1,\dots,t),\ \forall p \in \mathcal{P} \tag{3.27}$$

$$\sum_{t=1}^{T} z_{t,p} \leq \Gamma_p \qquad \forall p \in \mathcal{P} \tag{3.28}$$

$$0 \leq z_{t,p} \leq 1 \quad \forall t \in \mathcal{T},\ p \in \mathcal{P}$$

The objective function (3.26) in the adversarial sub-problem is a maximization problem. Constraint (3.27) is to ensure that the value of z becomes updated such that the optimal x value from the previous step, implementer sub-problem, remains feasible. Constraint (3.28) is for controlling the level of uncertainty in the model. Since in the implementer sub-problem z was a parameter, this constraint becomes redundant in the implementer sub-problem and is deleted from the set of constraints in the implementer sub-problem.

3.3.5 The Solution Algorithm

The following pseudocode presents the algorithm and the steps required to solve the problem and determine the optimal solution.

Algorithm 1 adversarial algorithm

$L = -\infty$, $U = +\infty$

$\mathcal{Z} = \{z_0\}$ and $z_0 = \{z_{t,p} | z_{t,p} = 0,\ \forall t \in T, \forall p \in P\}$

while $U \geq L$ **do**

SOLVE $\underset{x,c}{Min}\ \underset{z \in \mathcal{Z}}{Max}\ f(x, c, z)$

(x and c are decision variables and Solution is x^*,c^*)

UPDATE $L \leftarrow \underset{z \in \mathcal{Z}}{Max}\ f(x^*, c^*, z)$ while x^* and c^* are from previous step.

SOLVE $\underset{z}{Max}\ f(x^*, c^*, z)$ (Solution is z^*)

if U $\geq f(x^*, c^*, z^*)$ **then**

U=$f(x^*, c^*, z^*)$

end if

$\mathcal{Z} = \mathcal{Z} \cup z^*$

end while

Chapter 4

Numerical Results and Analysis

In this chapter, the proposed methodology is implemented on a set of numerical examples and the results are discussed. The model assumptions, variables, and parameters used in the numerical examples are explained in Section 4.1. Section 4.2 provides a sensitivity analysis on different cost components in the objective function. Finally, Section 4.3 presents a comparative analysis of the robust and deterministic models.

4.1 Model Assumptions, Variables, and Parameters

Parameters are as follows: The planning horizon, T, is 30 days. Four classes of priorities are considered in the model such that patients of priority one have the highest acuity level and patients of priority four have the lowest acuity level. The nominal demand $\bar{d}_t$ is generated based on a uniform distribution with a minimum of 100 and a maximum of 200. Daily demand is distributed among priorities as follows, 5% of demands are considered as patients with priority one, 20% of demands are priority two, 30% patients are priority three, and 45% patients are priority four. The possible deviation in the demand for each priority level is calculated by $\hat{d}_t = \frac{\bar{d}_t}{2}$. In this numerical analysis, the service time is different for different priority levels. The required service time for priorities 1, 2, 3, and 4 are considered as 3, 3, 2, and 1, respectively. Each priority level has a different wait-time target and the wait-time target considered for priorities 1, 2, 3, and 4 are 1, 2, 5, and 10, respectively. The budget of uncertainty Γ is defined for each priority level as follows: for the first priority 10, for the second priority 13.5, for the third priority 10, and for the fourth priority 6. These values are arbitrarily chosen; ideally for practical applications of this methodology, one must estimate these parameters based on real-world observations. As mentioned earlier, capacity is a decision variable in the model and a PWLC cost function is used for the capacity cost. In all the numerical examples in this chapter, the capacity cost function is a PWLC function shown in Figure 4.1. Table 4.1 summarizes the values of the model's parameters used for the numerical analysis.

Table 4.1: Parameters for Numerical Analysis

	Parameter	**value**			
	Time horizon	30			
	Number of priorities	4			
	Priority Levels	P_1	P_2	P_3	P_4
	Wait-time target	1	2	5	10
	Service level	3	3	2	1
	Budget of uncertainty (Γ_p)	10	13.5	10	6
	Distribution of demand across priorities	5%	20%	30%	45%
Demand	$\bar{d}$	$U(100, 200)$ [1]			
	$\hat{d}$	$\frac{1}{2}\bar{d}$			
Capacity	Ranges	0-200	200-250	250-380	
	Cost	0	3	10	

[1]Uniform Distribution

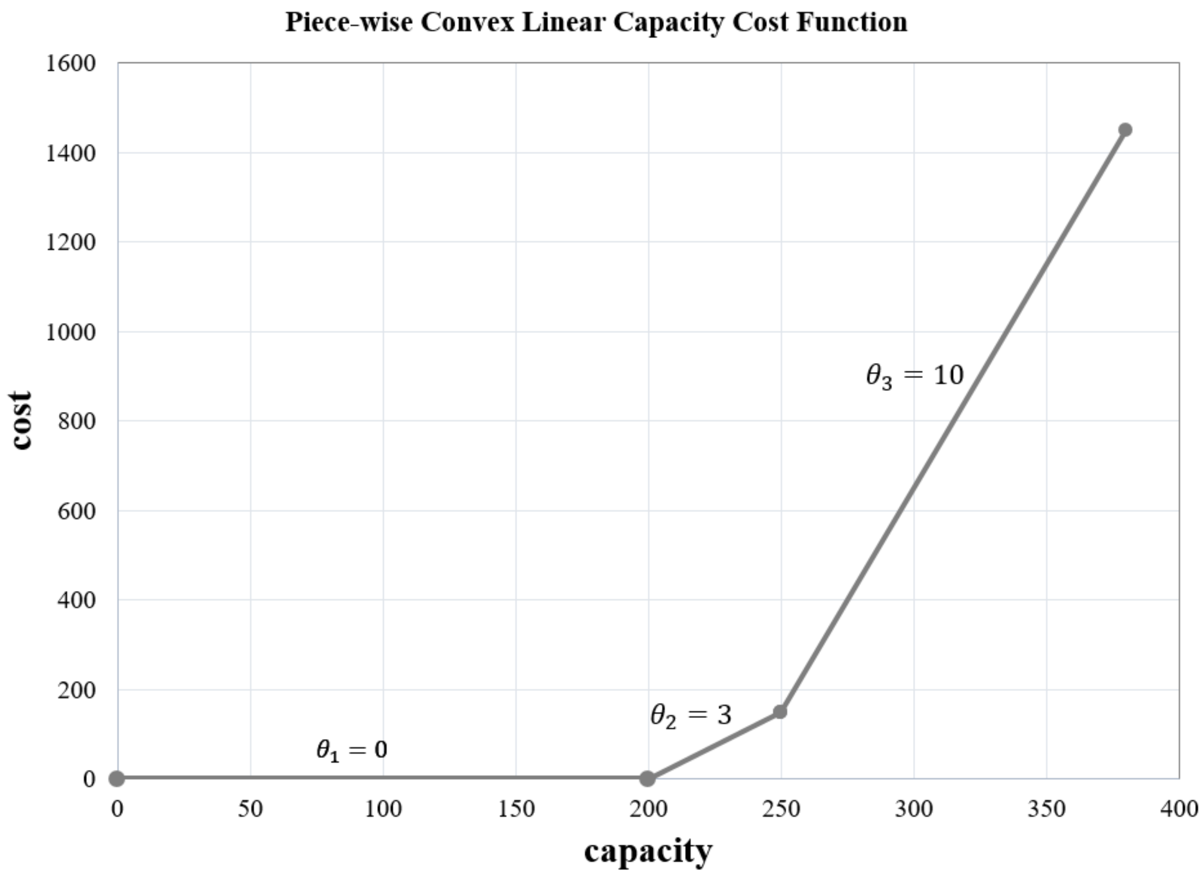


Figure 4.1: Piece-wise linear convex cost function considered in numerical examples

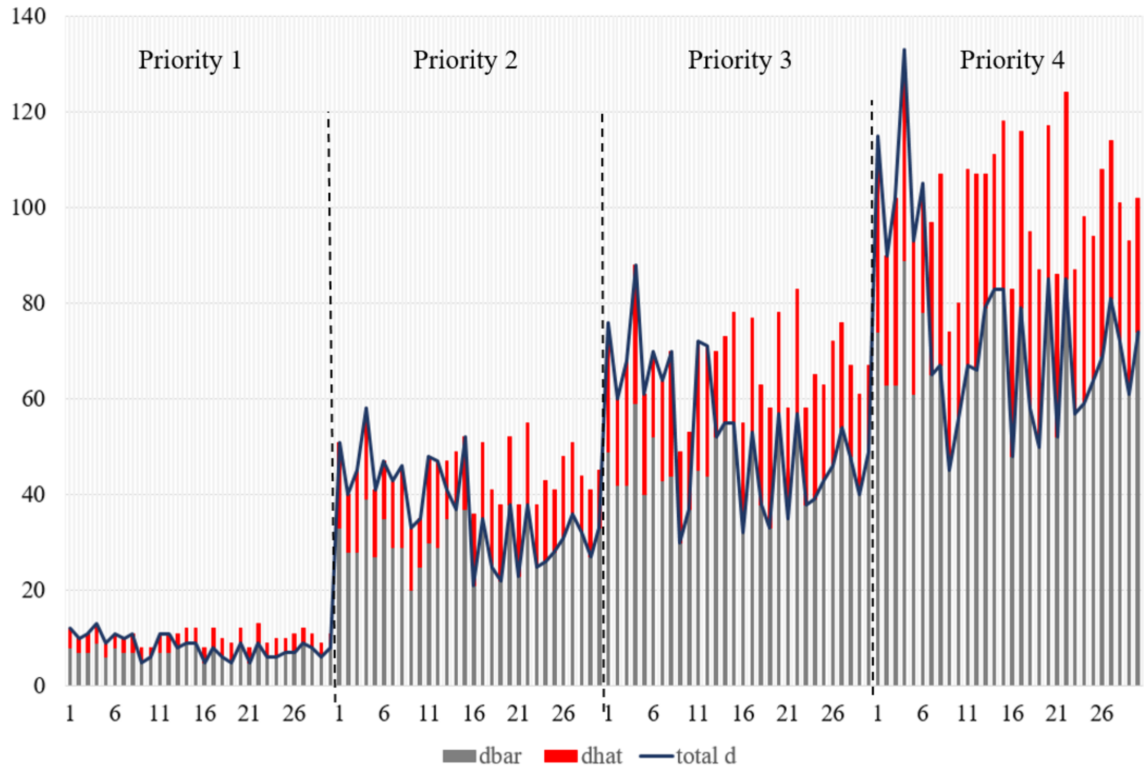
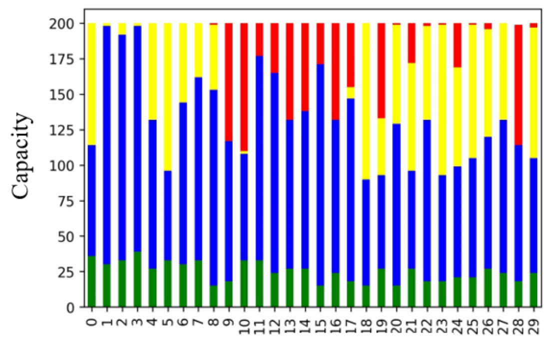


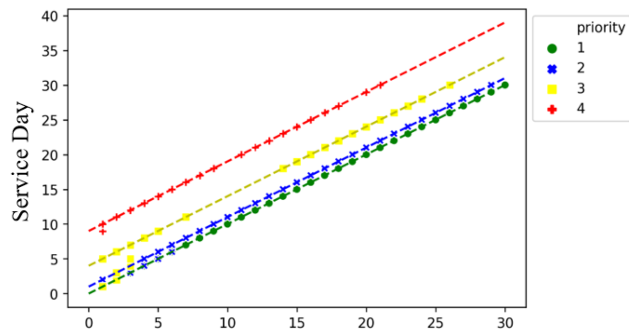
Figure 4.2: An example of $\bar{d}$ and $\hat{d}$ used for different priority levels

4.2 Sensitivity Analysis

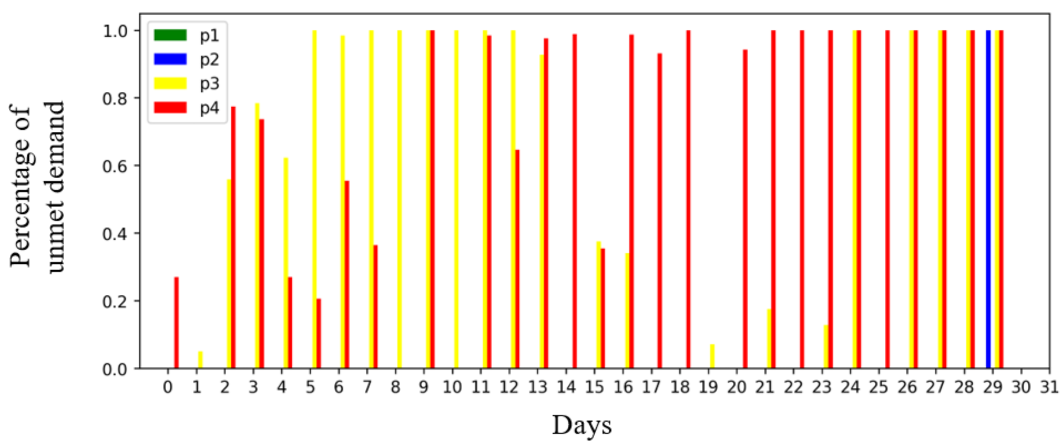
As mentioned before, the objective function has three cost components: the penalty cost for unmet demand (C_U), the capacity cost (C_C), and the penalty cost for late service (C_D). These three cost components are different in magnitude and given the current cost parameters in the model introduced in Section 3.2.3, C_C is larger than the other two cost components. Due to the imbalance in the magnitudes of the cost components, at optimality, if no normalization is used, the capacity level is kept at its minimum level during the horizon and C_U and C_D increase. Figure 4.3a shows the total capacity and the allocated capacity to each priority level on a daily basis in the horizon. Per Figure 4.3a, the amount of capacity assigned to patients of priority 4 is considerably insufficient and it leads to a large percentage of unmet demand of this priority, see Figure 4.3c. Figure 4.3b shows the arrival and appointment day of patients in a scatter-plot. The dashed lines in the plot show the allowed time window for each priority level, based on their arrival and their wait-time target. As illustrated, all the served patients have received service within their wait-time target. However, a large percentage of patients of priority 3 and 4 never received service within the time horizon (unmet demand due to insufficient capacity). The following section provides an analysis of how to normalize the three cost components in the objective function.



(a) Capacity allocation



(b) Arrival-day vs Service-day



(c) Percentage of unmet demand for different days and priorities when no normalization is used

Figure 4.3: coefficients of $((C_U), (C_C), (C_D))$= (1,1,1)

4.2.1 Capacity Cost Analysis

In the previous section, without cost normalization, at optimality, the daily capacity level is set at 200 for each day during the time horizon. In each day, a portion of the capacity is allocated to patients of priority 1 to satisfy the demand for this priority level. This allocation of capacity allows priority 1 patients to be served within their wait-time target. Similarly, all the patients of priority 2 who arrived between day 0 and day 28, received service within their wait-time target and only patients who arrived at the last day of time horizon do not receive service. The unmet demand at the end of the horizon for priority 2 is due to the end of the horizon effect and their wait-time target. On the contrary, for patients of priority 3 and 4, only a small portion of patients receive service during the time horizon and most of the patients do not receive service at all. This large amount of unmet demand of patients with priority 3 and 4 is due to the lack of capacity. As mentioned before, since the C_C is larger compare to the other cost components, the model avoids to increase the capacity level and decides to pay the penalty cost for unmet demand or late service time instead.

In this section, a sensitivity analysis of the different cost components in the objective function is conducted. The main goal of this analysis is to normalize the three cost components in the cost function. The coefficients a_1 and a_2 are incorporated in the model such that a_1 is the coefficient of $C_D + C_U$ and a_2 is the coefficient of C_C. Thus, the objective function becomes:

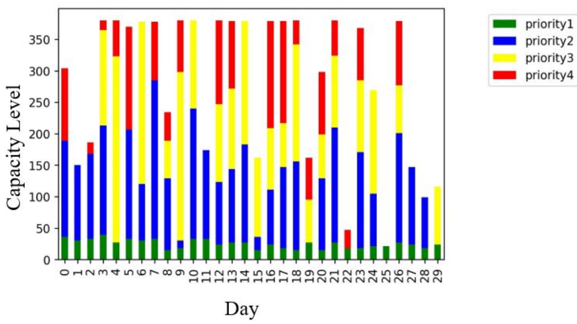
$$a_1 \times \big(C_D + C_U\big) + a_2 \times C_C$$

such that,

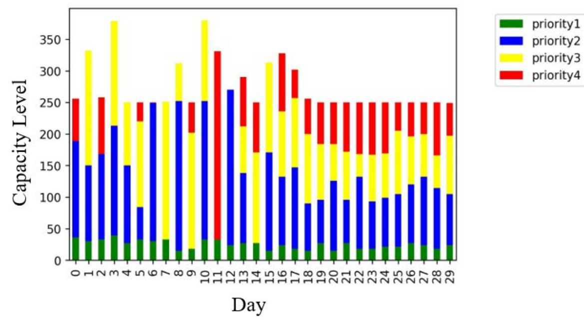
$$a_1 + a_2 = 1.$$

Different values for a_1 and a_2 are tested and the results are presented in Figures 4.4 to 4.6. In figures 4.4, 4.5, and 4.6 capacity allocation, arrival/service time, and ratio of unmet demand are illustrated for different capacity cost coefficients, a_2, ranging from 0.1 to 0.45. For the sake of brevity, only the results from a few selected scenarios are presented here.

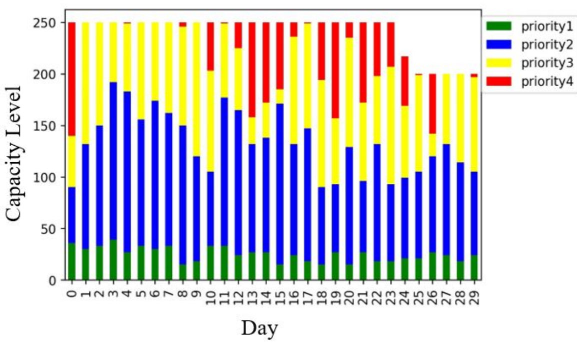
Figure 4.4a represents the case with no capacity cost ($a_2 = 0$). Figure 4.4a shows the required level of capacity to ensure that all the patients whose allowable time window falls within the time horizon receive service.



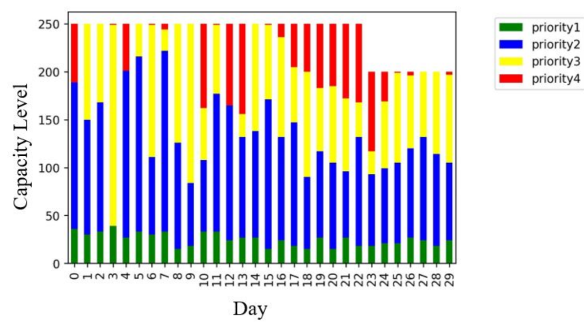
(a) $(a_2,\ a_1)$=(0,1)



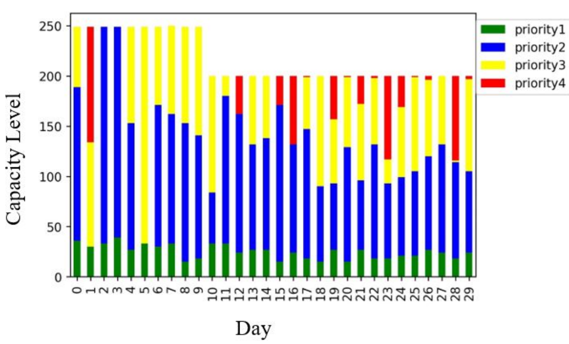
(b) (a_2, a_1)=(0.1,0.9)



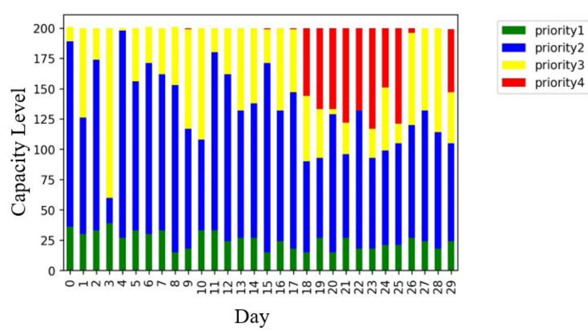
(c) (a_2, a_1)=(0.15,0.85)



(d) (a_2, a_1)=(0.2,0.8)

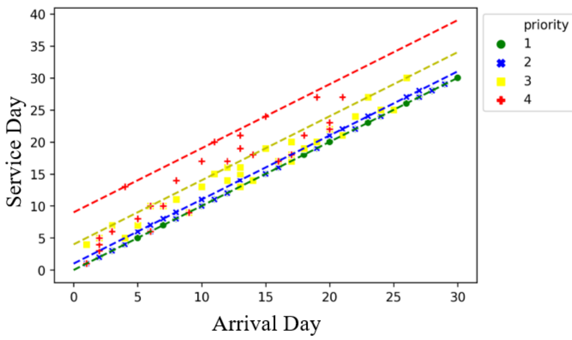


(e) (a_2, a_1)=(0.4,0.6)

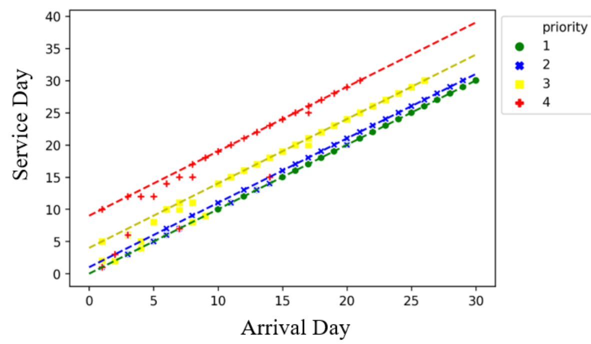


(f) (a_2, a_1)=(0.45,0.55)

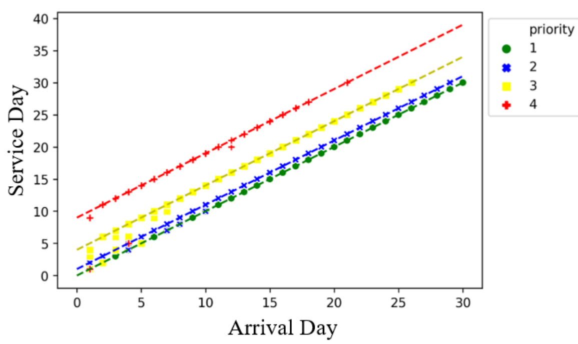
Figure 4.4: Capacity Allocation for different values of (a_2, a_1)



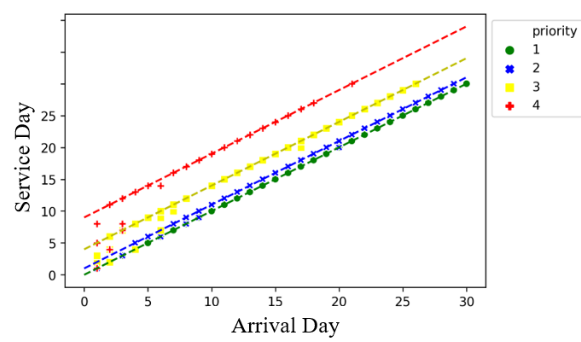
(a) $(a_2,\ a_1)$=(0,1)



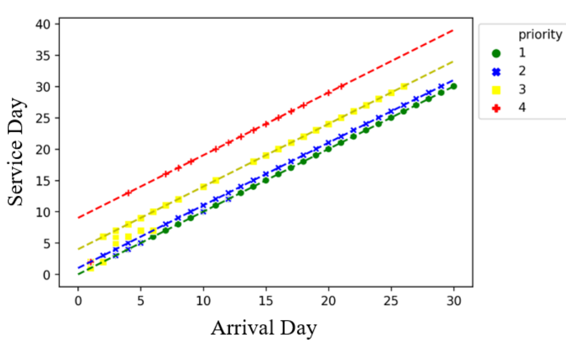
(b) (a_2, a_1)=(0.1,0.9)



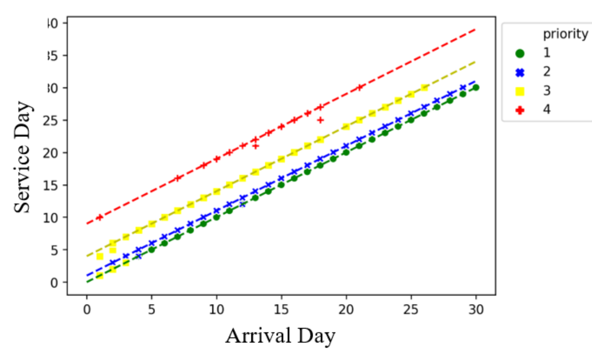
(c) (a_2, a_1)=(0.15,0.85)



(d) (a_2, a_1)=(0.2,0.8)



(e) (a_2, a_1)=(0.4,0.6)

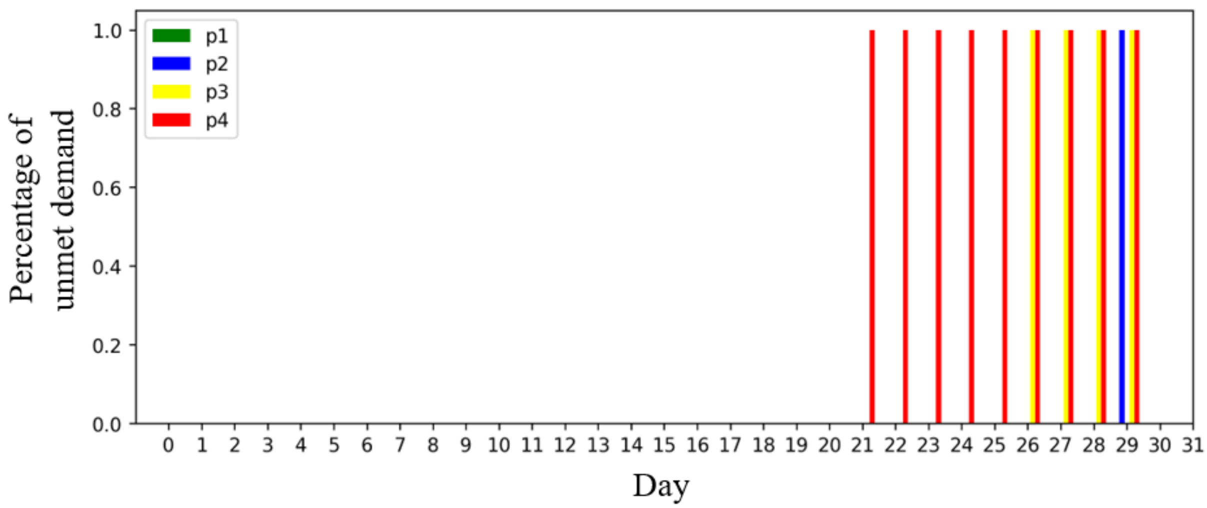


(f) (a_2, a_1)=(0.45,0.55)

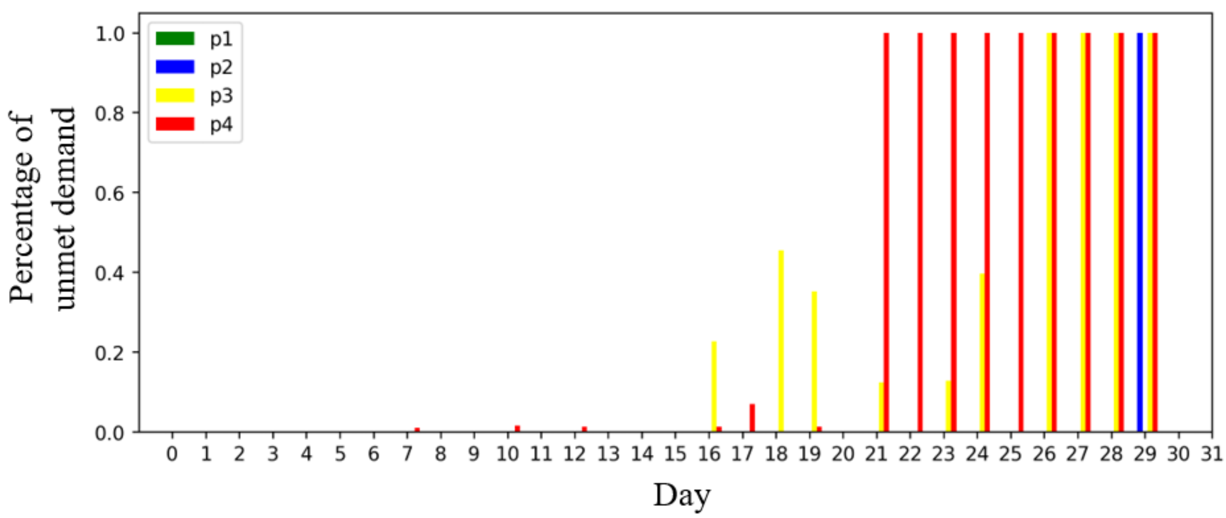
Figure 4.5: Scatter-plot of arrival day vs. service day for different values of (a_2, a_1)

Figure 4.6 shows the percentage of patients who do not receive service within the time horizon. As can be seen in Figure 4.6a, all patients of priority 1 receive service. However, all the patients of priority 4 who arrived after day 21, do not receive service. This stems from the fact that the allowable time window for patients of priority 4 who arrive after day 21, goes beyond the time horizon. The allowable time window for each patient refers to the time interval in which that patient can receive service without any late service penalty cost ($C_D = 0$). For a patient of priority p with wait-time target w_p who arrives at day t, the allowable time window is $(t, t + w_p - 1)$. The same interpretation holds true for the unmet demand of patients with priority 2 and 3 near the end of the horizon. Comparing Figures 4.6a, 4.6b, 4.6c, 4.6d, 4.6e and 4.6f, it can be seen that once a_2 goes beyond 0.2, the amount of unmet demand increases drastically. Thus, the capacity cost coefficient for the rest of the analysis is set at 0.15 which leads to a more reasonable percentage of unmet demand during the horizon.

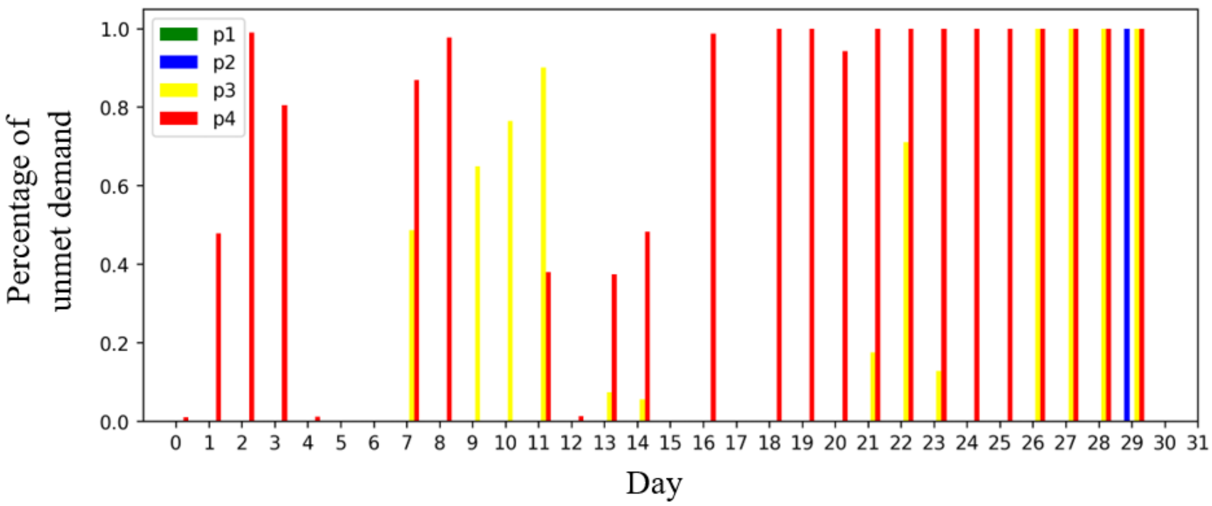
In Section 4.2.2, in addition to the results for $a_2 = 0.15$, the results for $a_2 = 0.4$ are also provided. The reason behind considering $a_2 = 0.4$ in this analysis is to determine the impact of penalty cost for unmet demand coefficient, a'_1, and penalty cost for delayed patients coefficients, a'_3, on the model under "a more expensive capacity".



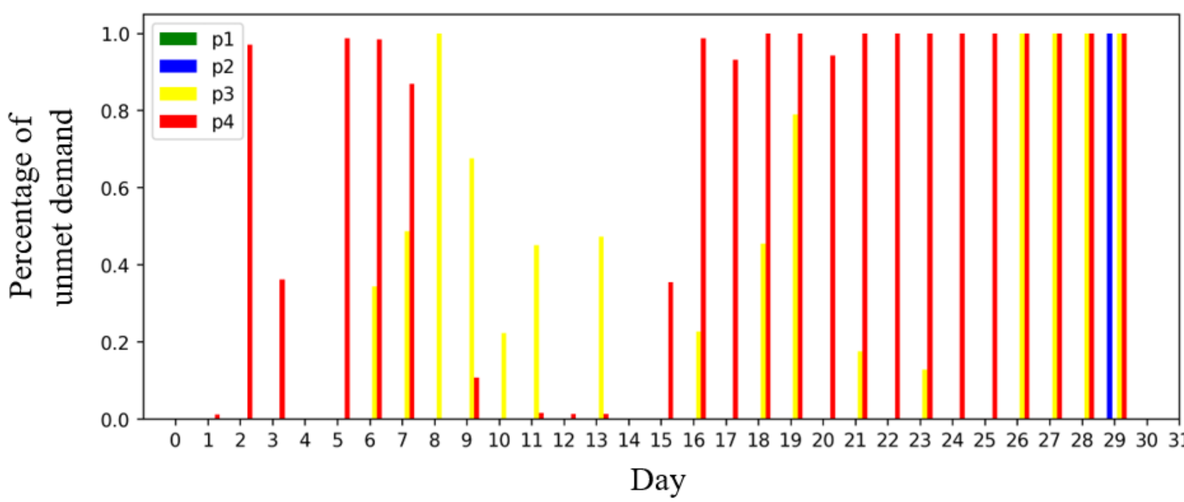
(a) (a_2, a_1)=(0,1)



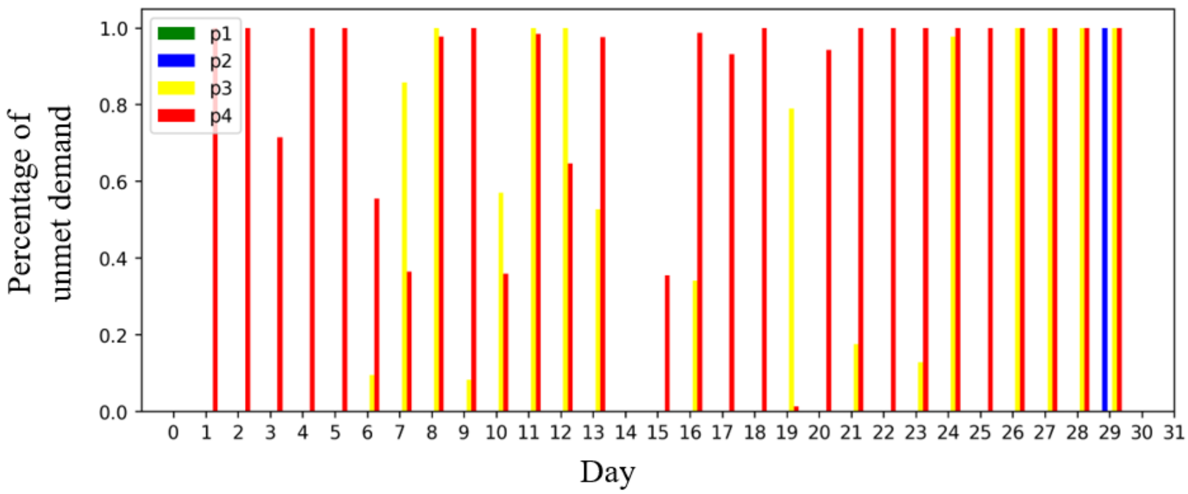
(b) (a_2, a_1)=(0.1,0.9)



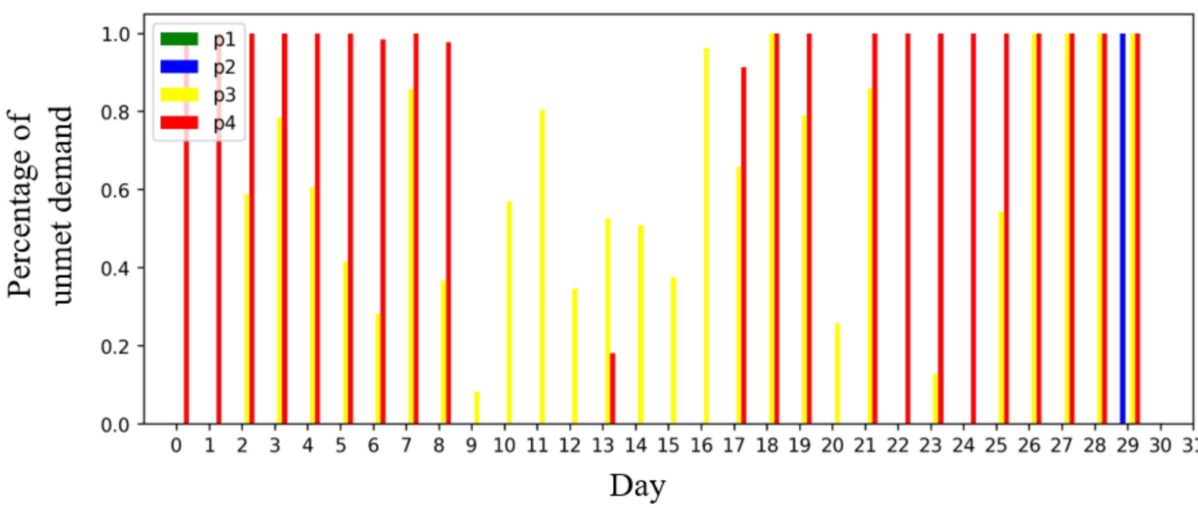
(c) (a_2, a_1)=(0.15,0.85)



(d) (a_2, a_1)=(0.2,0.8)



(e) (a_2, a_1)=(0.4,0.6)



(f) (a_2, a_1)=(0.45,0.55)

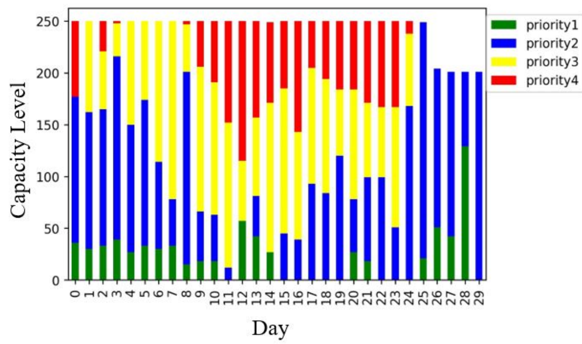
Figure 4.6: Percentage of unmet-demand during the horizon for different values of (a_2, a_1)

4.2.2 Delayed vs. Unmet Demands Penalty Coefficients Analysis

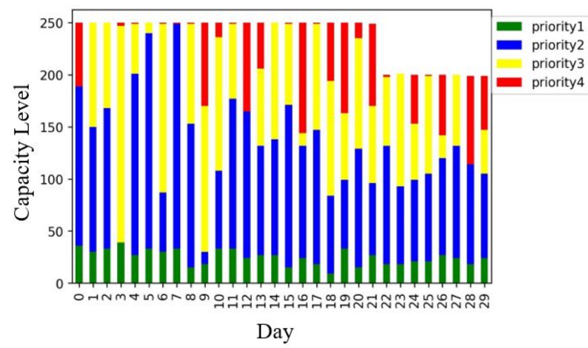
In this section, the impact of C_U and C_D are studied while the capacity cost coefficient, a_2, is fixed. The results are provided for two different values of a_2, 0.15 and 0.4. The two values of a_2 are selected in accordance with the plausible range for the capacity cost coefficient as identified in Section 4.2.1. For values less than $a_2 = 0.15$, the model increases the capacity to its maximum allowed value which is not desirable from a practical point of view. For values of $a_2 \geq 0.4$, the model perceives the additional capacity too costly to the extent that it does not consider any additional capacity which is again not logical from a practical point of view. The coefficients, a'_1 and a'_3 are associated with the C_U and C_D, respectively, such that $a'_1 + a_2 + a'_3 = 1$.

First Scenario: Capacity Cost Coefficient (a_2) = 0.15

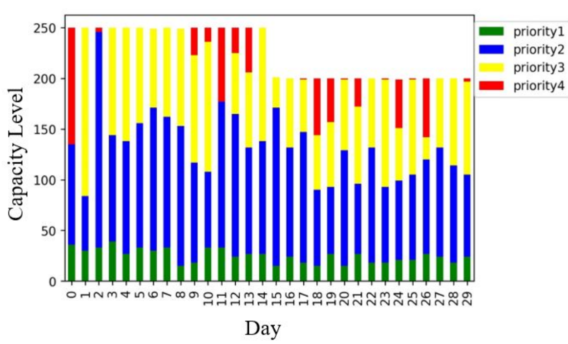
In this section, different combinations of a'_1 and a'_3 are tested for the fixed capacity cost coefficient of $a_2 = 0.15$. Figures 4.7, 4.8, 4.9, and 4.10 show the capacity allocation, unmet demand, and the arrival/service time for four different set of cost coefficients for C_U and C_D. Moving from Figures 4.7a to 4.7d, 4.8a to 4.8d, 4.9a to 4.9d, and 4.10a to 4.10d the coefficient for C_D, a'_3, increases and on the other hand the coefficient for C_U, a'_1, decreases.



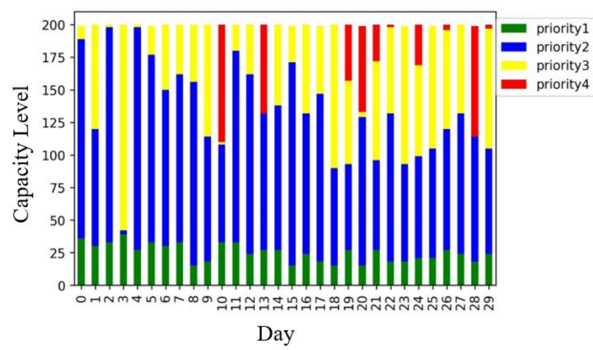
(a) (0.8,0.15,0.05)



(b) (0.55,0.15,0.3)

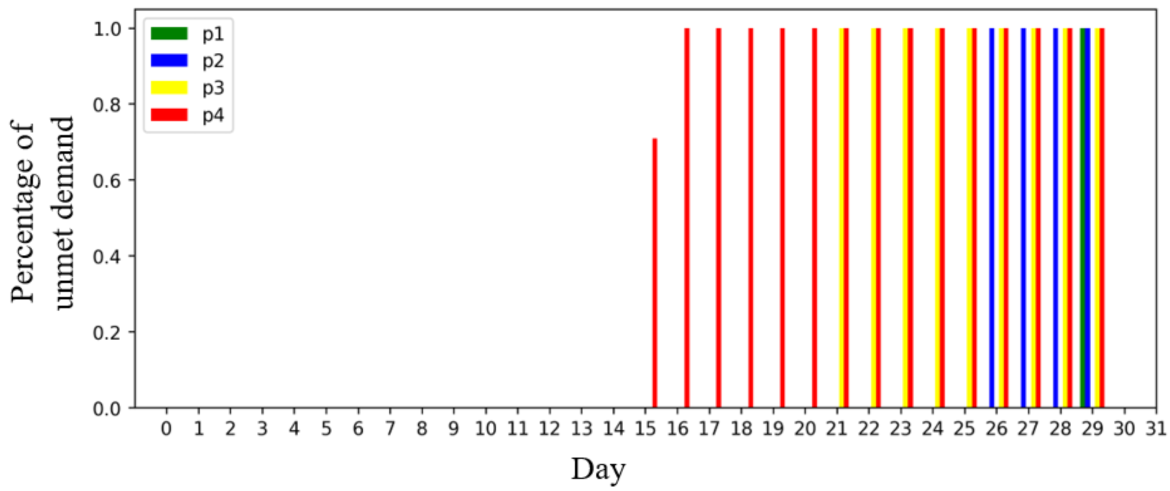


(c) (0.3,0.15,0.55)

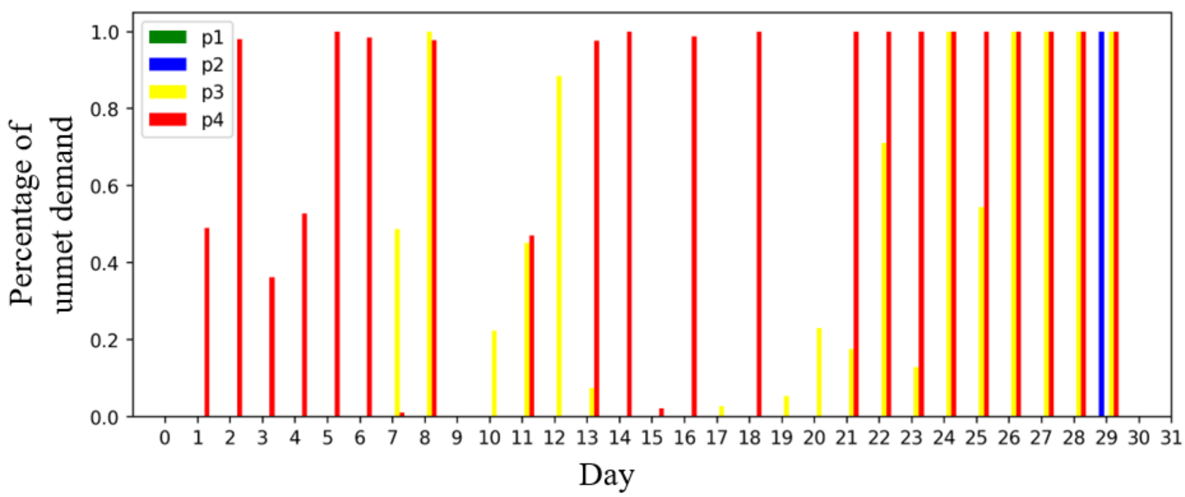


(d) (0.05,0.15,0.8)

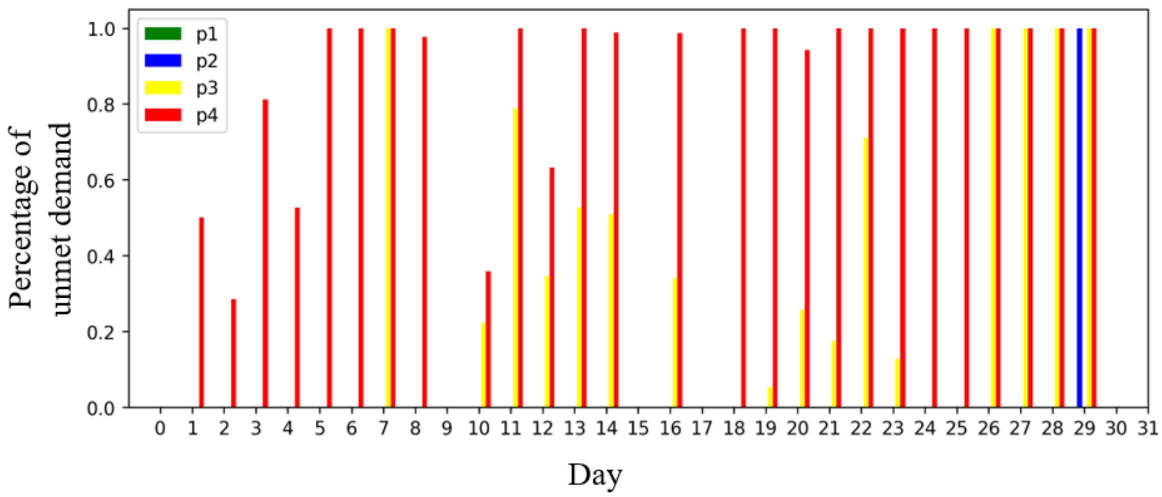
Figure 4.7: Capacity Allocation



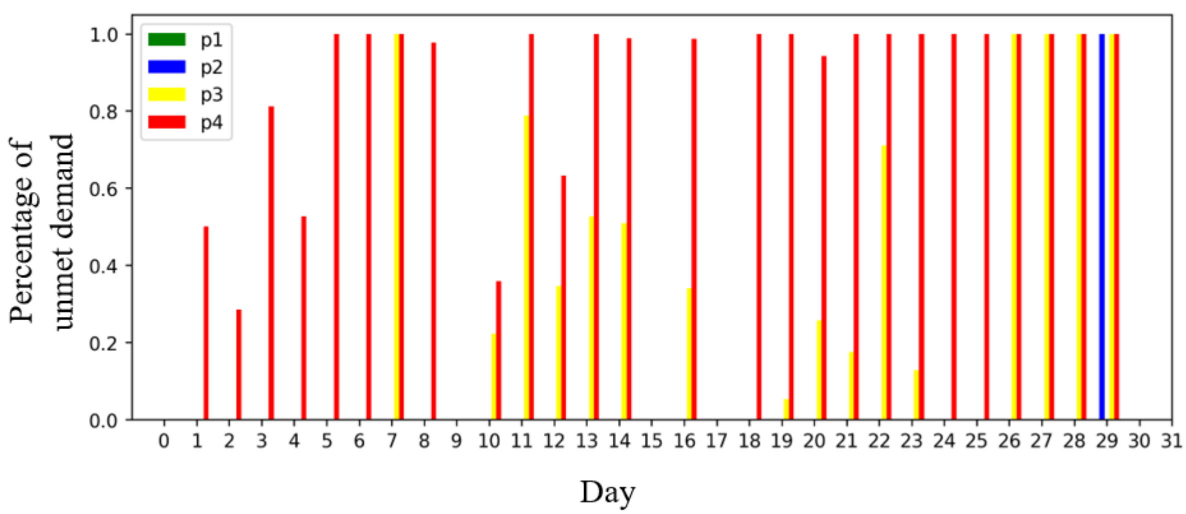
(a) (a'_1, a_2, a'_3)=(0.8,0.15,0.05)



(b) (a'_1, a_2, a'_3)=(0.55,0.15,0.3)

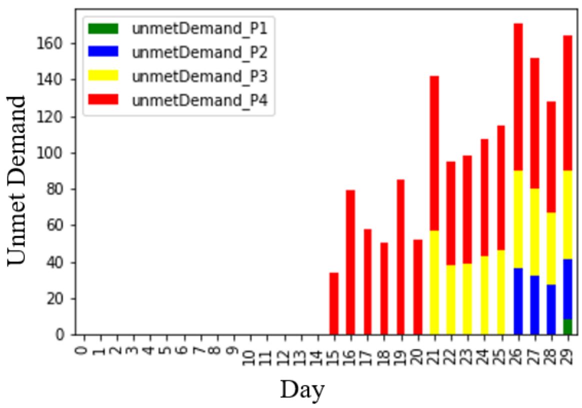


(c) (a'_1, a_2, a'_3)=(0.30,0.15,0.55)

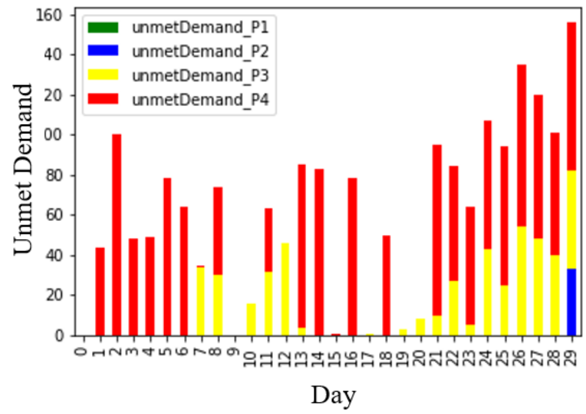


(d) (a'_1, a_2, a'_3)=(0.05,0.15,0.8)

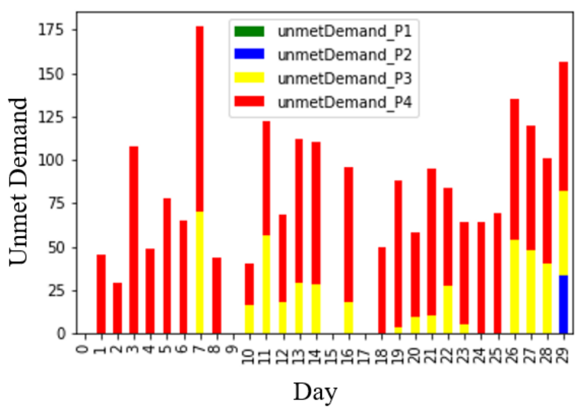
Figure 4.8: Percentage of Unmet Demand



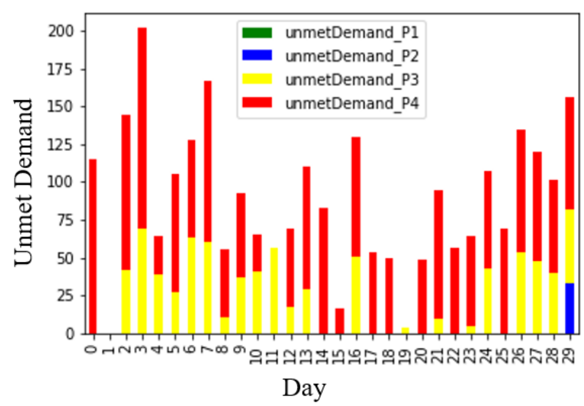
(a) (0.8,0.15,0.05)



(b) (0.55,0.15,0.3)



(c) (0.3,0.15,0.55)

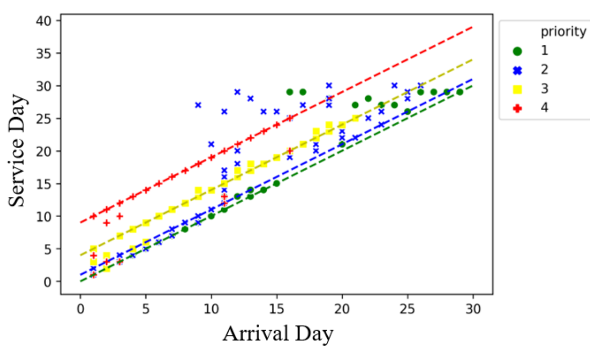


(d) (0.05,0.15,0.8)

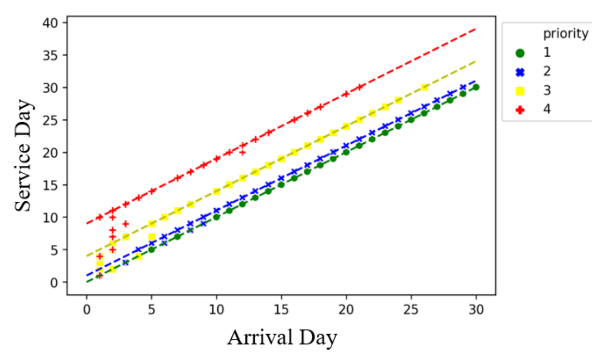
Figure 4.9: Unmet Demand

Figure 4.10 illustrates that by increasing the coefficient of C_D, the number of patients who are served after their wait-time target (delayed patients) decreases drastically. Comparing Figures 4.10a and 4.10b, when the coefficient of $C_U = 0.8$, some patients wait beyond their wait-time target, but when coefficient of $C_U = 0.55$, the wait times of all the patients who receive service fall within their wait-time target. Hence, the model would rather have unmet demand instead of delayed patients. Even though a_2 remains unchanged in all the provided examples, the average level of capacity for the horizon decreases as a result of an increase in the coefficient of delayed patients penalty cost and a decrease in the coefficient of unmet demand penalty cost.

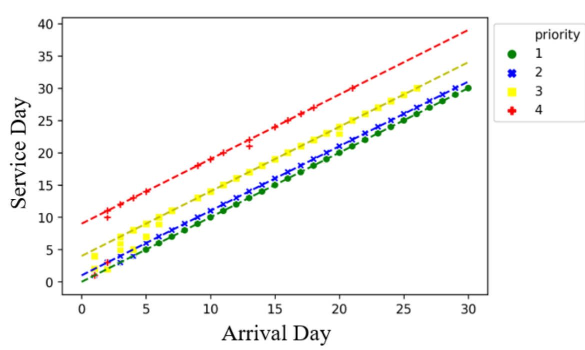
A sensitivity analysis of a_1' is conducted. It reveals that for the range of 0.66 to 0.80 for the unmet demand penalty cost coefficient when the capacity cost coefficient is 0.15, some patients receive service with a delay.



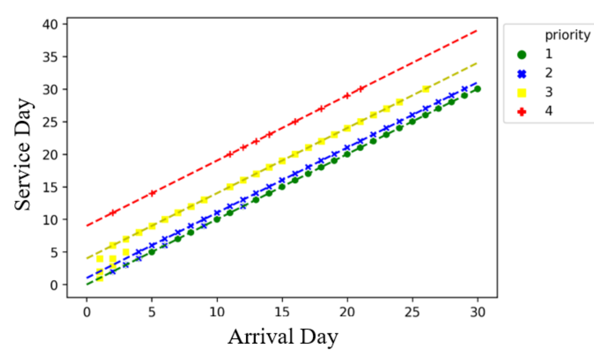
(a) (0.8,0.15,0.05)



(b) (0.55,0.15,0.3)



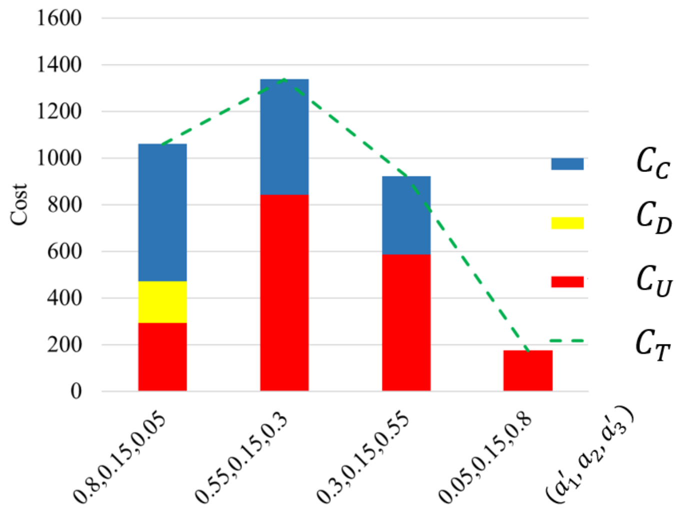
(c) (0.3,0.15,0.55)



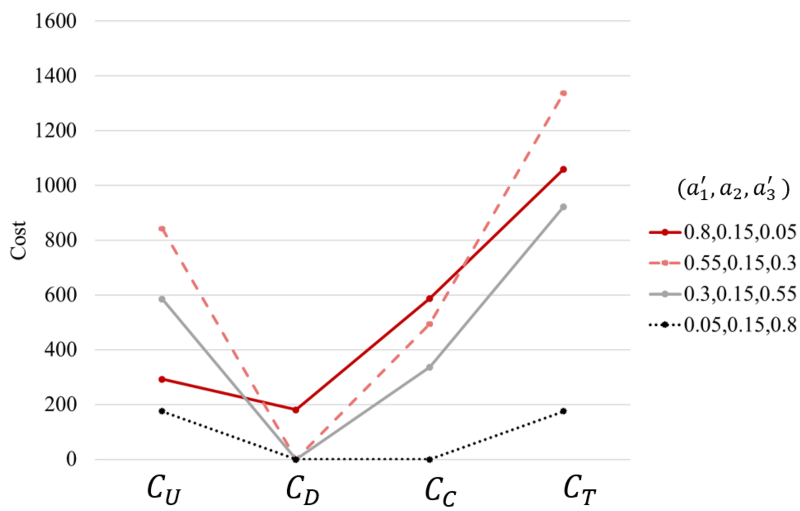
(d) (0.05,0.15,0.8)

Figure 4.10: Scatter-plot of Arrival Day vs. Service Day

Figure 4.11 illustrates the values of C_C, C_D, C_U, and C_T for different values of cost coefficients a'_1, a_2, and a'_3. The changes in a'_1 and a'_3 affect the C_C, despite the fact that the coefficient for C_C remains constant at 0.15. C_C decreases with a decrease in a'_1 to the extend that it reaches 0 at $a'_1 = 0.05$. Based on the tested values for a'_1 in Figure 4.11 the threshold that C_D becomes insensitive to changes in a'_3 is at 0.05.



(a) Objective Cost details ($a_2 = 0.15$)

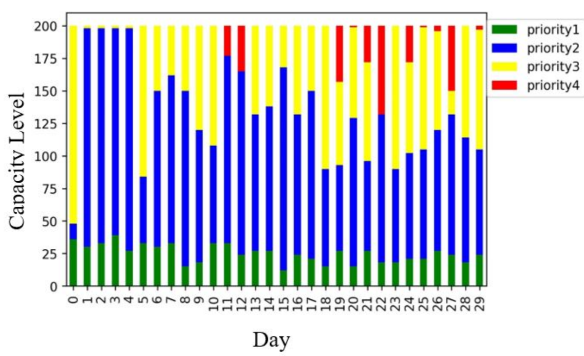


(b) Objective Cost details ($a_2 = 0.15$)

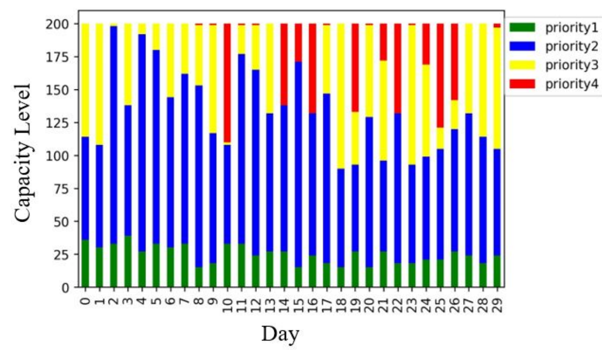
Figure 4.11: Cost function details at a_2 =0.15

Second Scenario: capacity cost coefficient (a_2)= 0.4

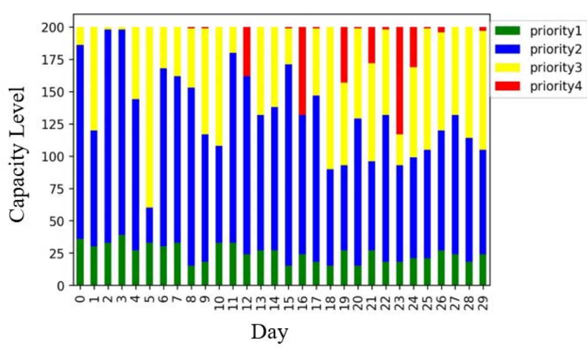
This section presents the results of the sensitivity analysis on the coefficients of C_D and C_U when $a_2 =$ 0.4. Figure 4.12 shows the capacity allocation for different combinations of a'_1 and a'_3. The total capacity level across all scenarios is 200, which is at the maximum capacity level in the PWLC capacity cost function with no cost and is the result of the larger coefficient, 0.4, for capacity cost.



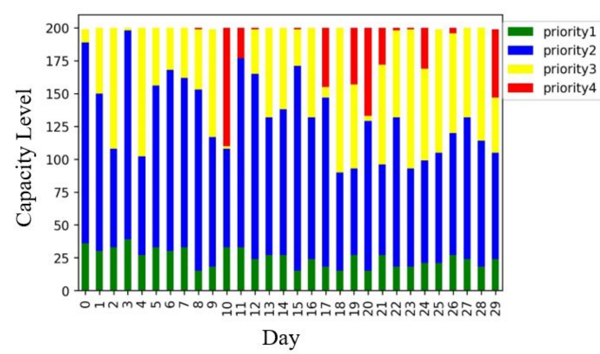
(a) (0.45,0.4,0.15)



(b) (0.3,0.4,0.3)



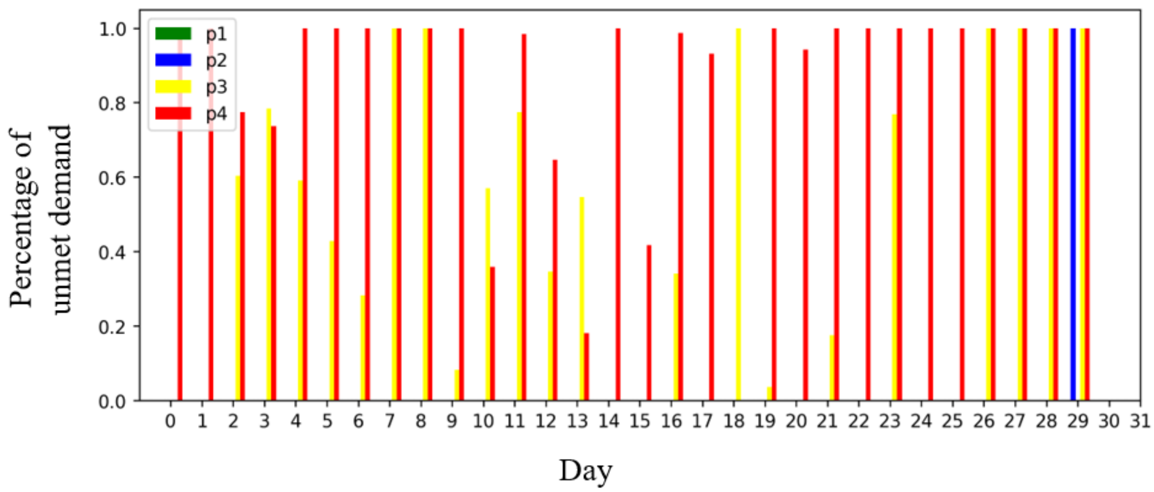
(c) (0.15,0.4,0.45)



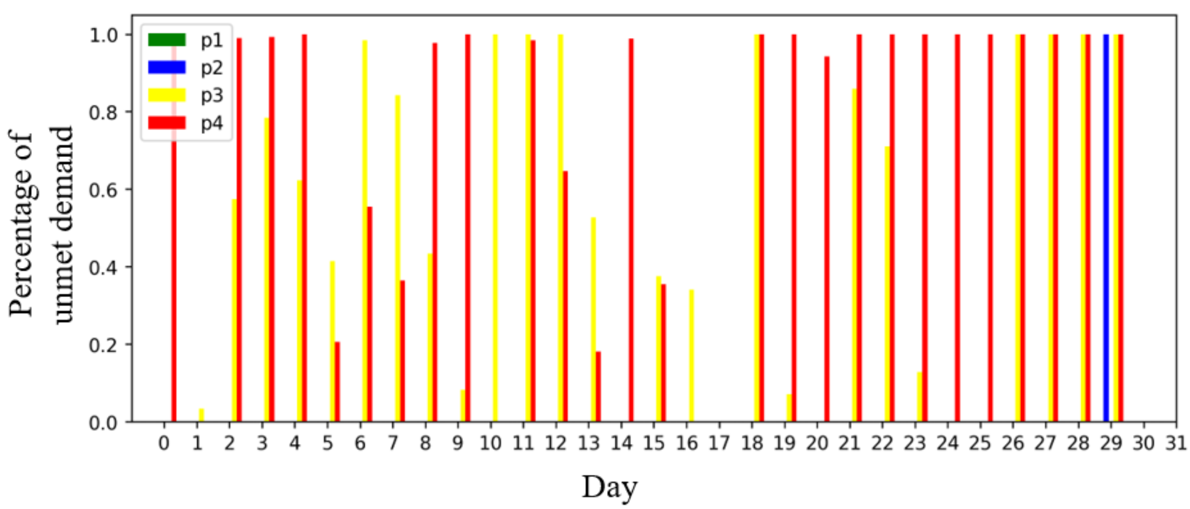
(d) (0.05,0.4,0.55)

Figure 4.12: Capacity Allocation

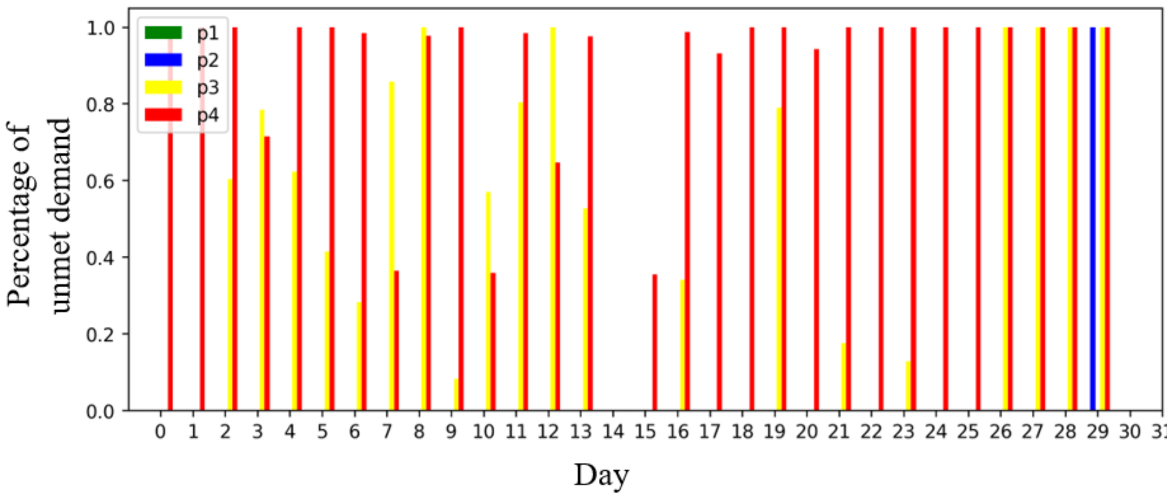
Figure 4.13 denotes the percentage of unmet demand for different combinations of a'_1 and a'_3, where $a_2 =$ 0.4.



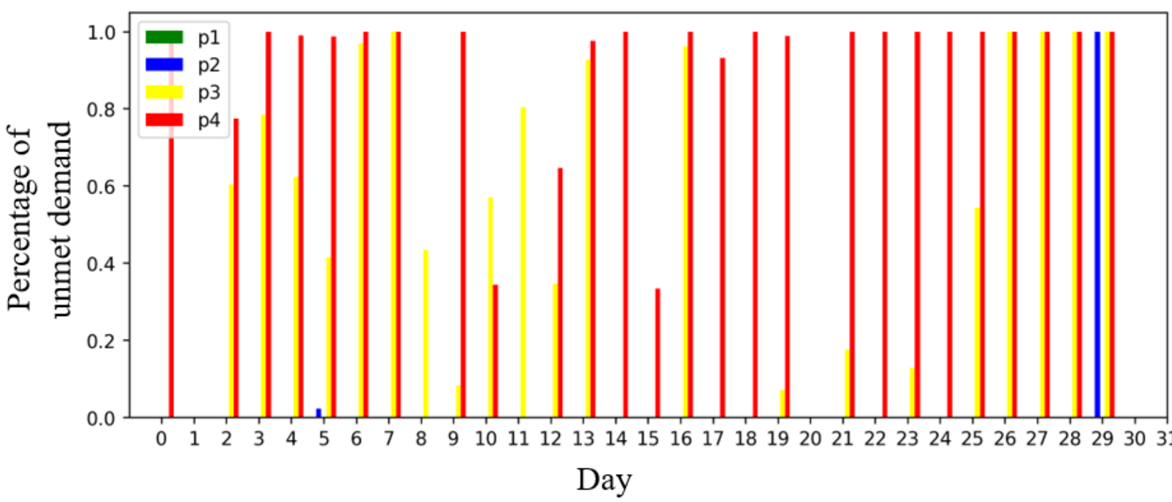
(a) (a'_1, a_2, a'_3)=(0.45,0.4,0.15)



(b) (a'_1, a_2, a'_3)=(0.3,0.4,0.3)

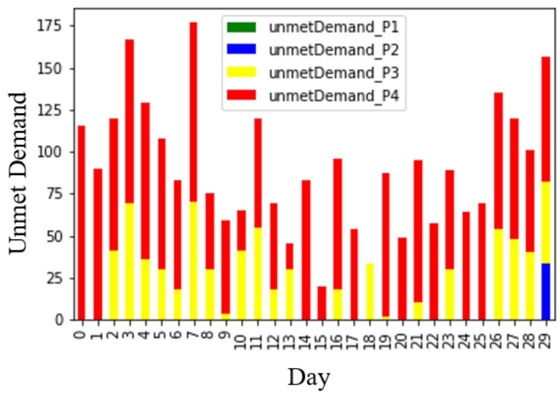


(c) (a'_1, a_2, a'_3)=(0.15,0.4,0.45)

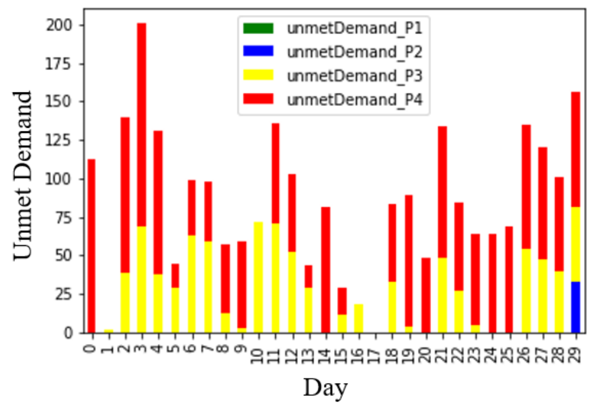


(d) (a'_1, a_2, a'_3)=(0.05,0.4,0.55)

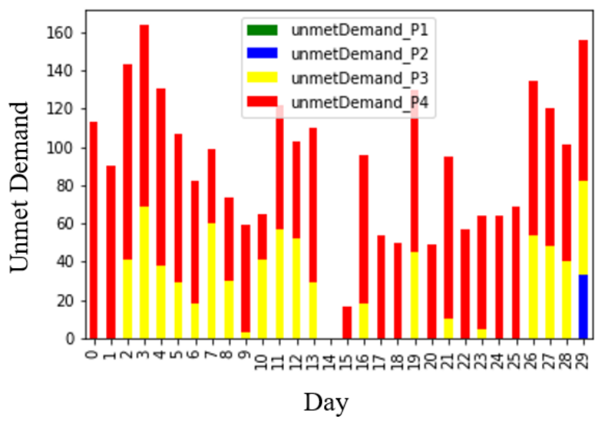
Figure 4.13: Percentage of unmet demand



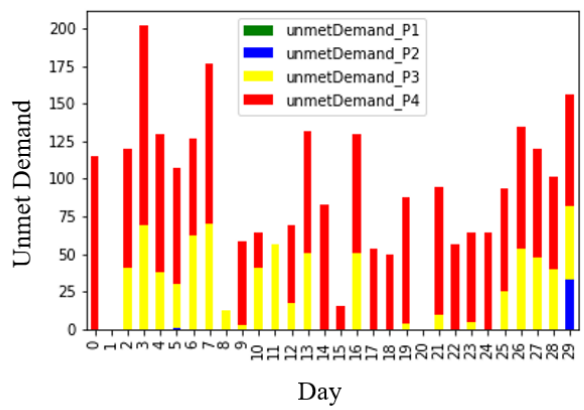
(a) (0.45,0.4,0.15)



(b) (0.3,0.4,0.3)



(c) (0.15,0.4,0.45)

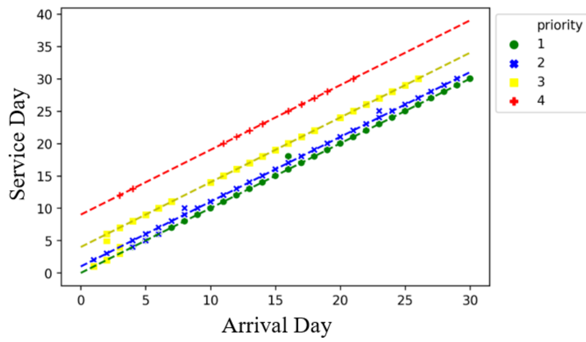


(d) (0.05,0.4,0.55)

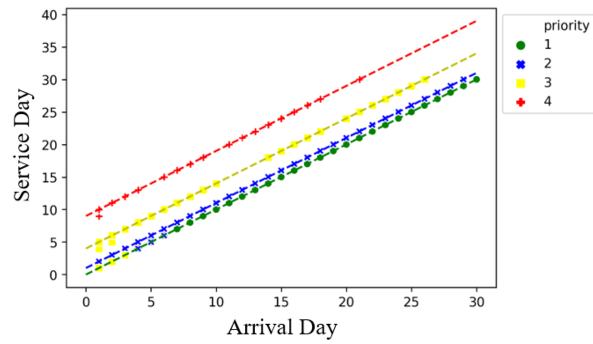
Figure 4.14: Unmet Demand

Figure 4.15, similar to Figure 4.10, illustrates the arrival and service dates and allowable time window for each priority. The scatter-plots representing the arrival and service days for each priority fall below the solid line which represents the wait-time target, meaning that all patients have received service within their wait-time target. As can be seen in Figure 4.16b, the only contributing element to cost function across all scenarios is C_U and

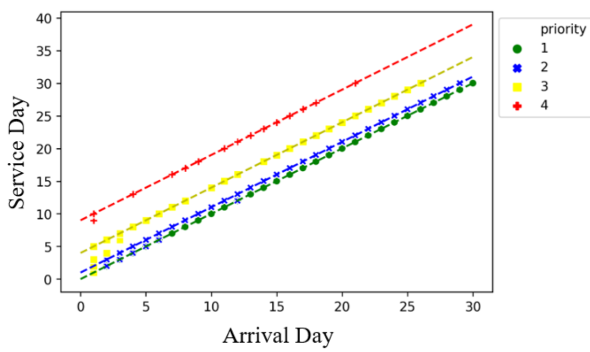
C_D; and C_C is zero across all scenarios.



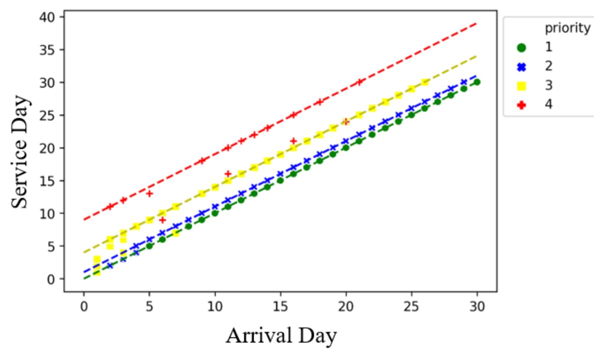
(a) (0.45,0.4,0.15)



(b) (0.3,0.4,0.3)



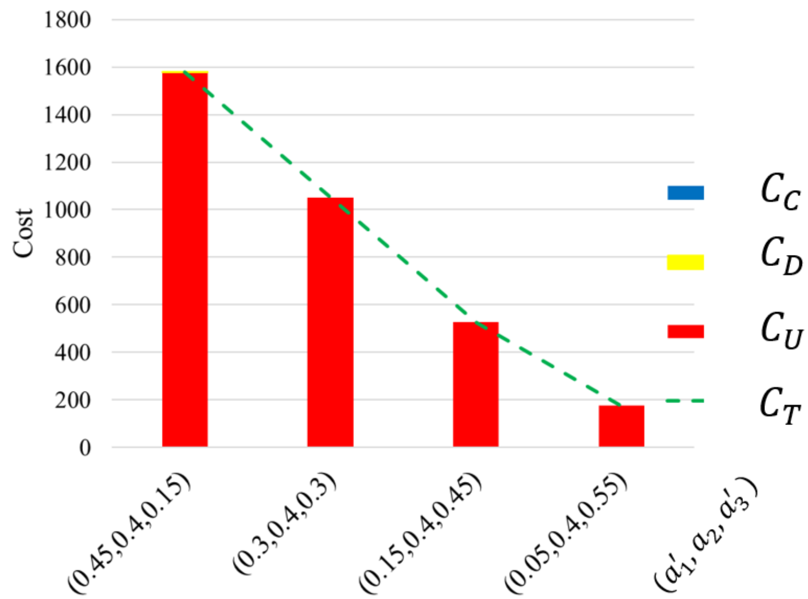
(c) (0.15,0.4,0.45)



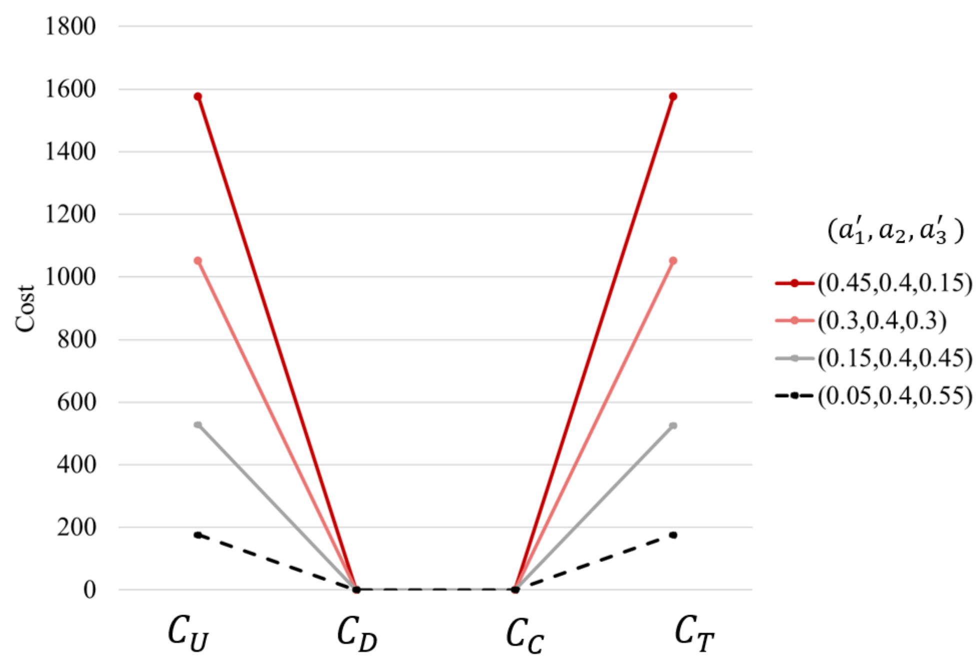
(d) (0.05,0.4,0.55)

Figure 4.15: Scatter-plot of Arrival vs. Service day

Figure 4.16 illustrates break-down of the cost function for each scenario; C_C is zero across all scenarios meaning that the model tries to keep the capacity level at minimum to avoid large capacity costs. Figure 4.16 shows that the total cost decreases with the decrease in C_U; it is noted that C_C without considering the coefficients are of the same order (around 3500) and the decreasing trend is merely a reflection of the decreasing a'_1.



(a) Objective cost details ($a_2 = 0.4$)



(b) Objective cost details ($a_2 = 0.4$)

Figure 4.16: Cost function details at a_2= 0.4

This analysis clearly shows that if the capacity cost coefficient, a_2, or capacity cost values (piece-wise linear capacity cost function) are set too high, the model would not be able to strike a balance between the delayed/unmet patients penalty costs and the capacity costs which would lead to non-optimal results and misleading outcomes. It is concluded that the modelling results are highly sensitive to the cost function coefficients and the capacity cost function. It is then necessary to use calibration techniques or expert opinion to determine the cost function coefficients and capacity cost levels. For the numerical analysis of the model in the next section, the costs for delayed and unmet demand are combined, which is more intuitive as the unmet demand is nothing but a delayed demand that receives service after the end of the horizon. According to the sensitivity analysis in Section 4.2.1, the coefficients for capacity cost, a_2, is set to 0.15 and the coefficient for unmet/delayed patients penalty cost, a_1, is set to 0.85.

4.3 Robust vs. Deterministic

In this section, the proposed robust model and deterministic model are compared in terms of the quality of their solutions under different realizations of demand, the application of either of models in planning purposes, and their cost functions.

4.3.1 Performance Analysis

In this section, the optimal solution of the robust model and the deterministic model are compared by using some performance measures such as average waiting time and service level for each priority level. Figure 4.17 shows the steps for comparing the deterministic model and robust model. First, the nominal demand and the possible deviations, $\bar{d}$ and $\hat{d}$, are inputted into both the deterministic and robust models and capacity allocations for the inputted demand are extracted from each model. Next, the performances of the capacity allocation of each model are evaluated under 100 random demands within the budget of uncertainty in terms of cost, average wait time, and service level. Hereafter, I refer to these 100 random demands as scenarios.

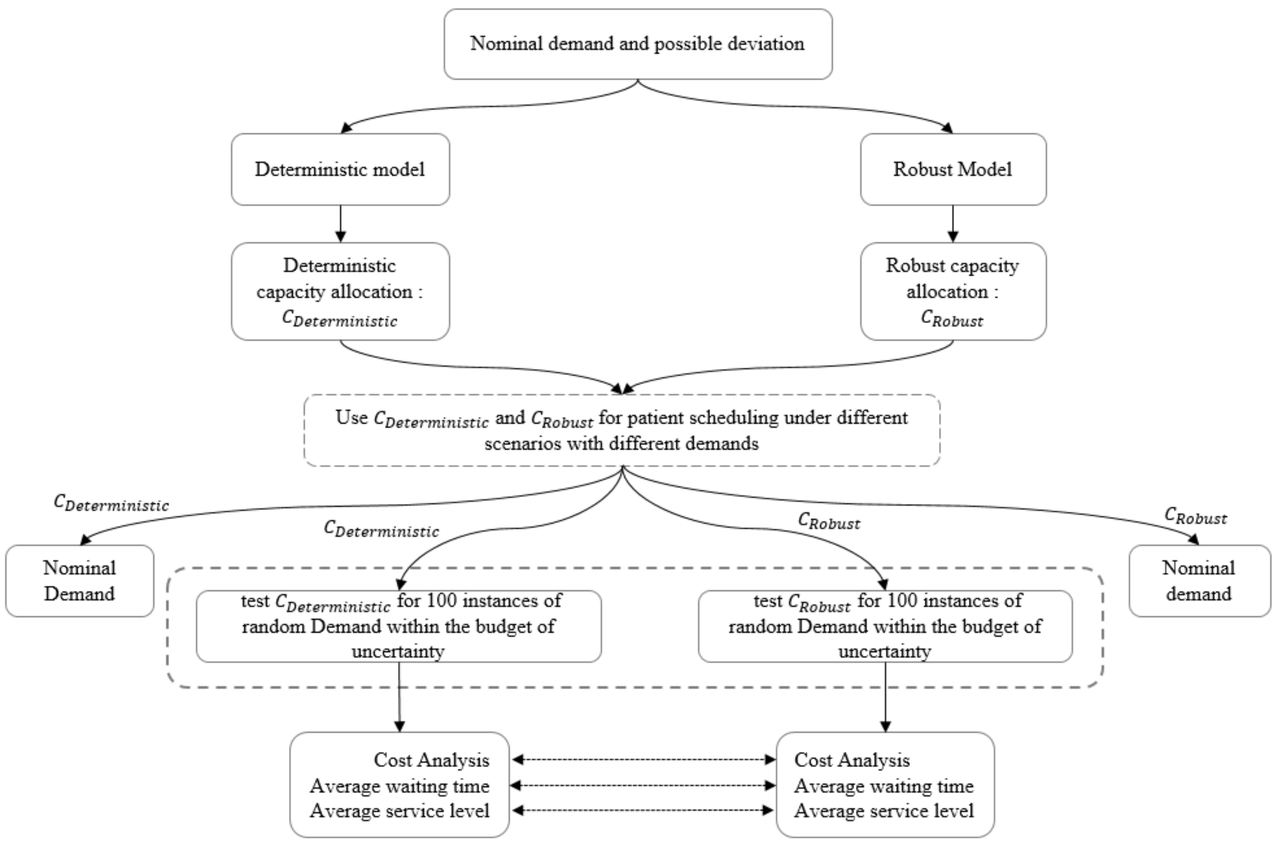


Figure 4.17: Flowchart showing steps for Robust vs. Deterministic Analysis

The stacked bars in Figure 4.18 show the nominal demands and deviations used in all the scenarios. The blue line on the chart shows an instance of random demand within the budget of uncertainty.

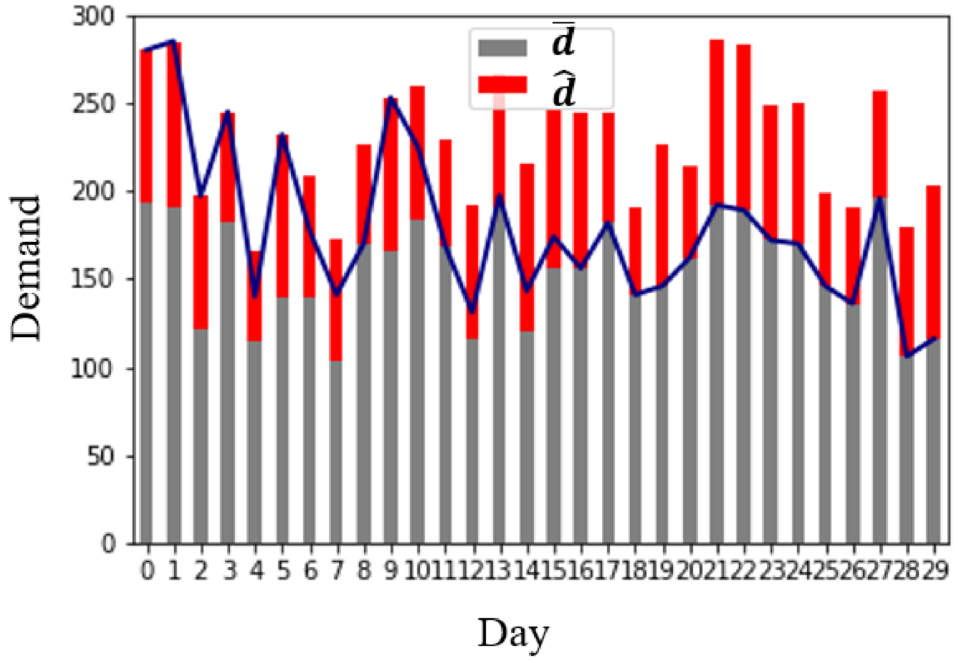
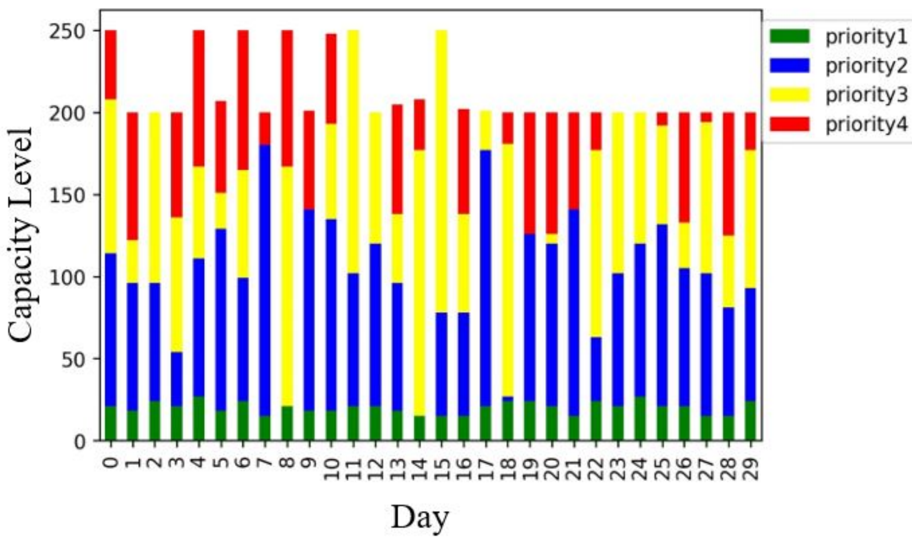
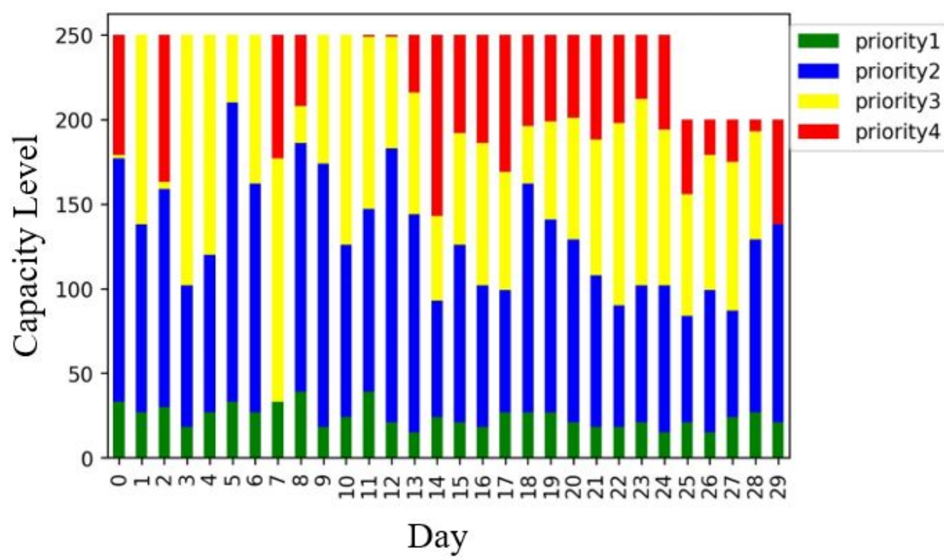


Figure 4.18: The nominal demand and the possible deviation used for robust vs. deterministic analysis

Figure 4.19a and 4.19b show the capacity allocation of the deterministic and the robust models for each priority, respectively. Comparing Figure 4.19a and Figure 4.19b, the average capacity level in the deterministic model is less than the capacity level in the robust model. This lower level of capacity in the deterministic model stems from the fact that the demand in the deterministic model is equal to the nominal demand, $\bar{d}$, which is shown in Figure 4.18 and does not include the demand uncertainty.



(a) $C_{Deterministic}$



(b) C_{Robust}

Figure 4.19: Capacity Allocation of Deterministic and Robust Model ($C_{Deterministic}$ and C_{Robust})

Figure 4.20 shows the cost analysis of deterministic and robust models. Note that, normalized versions of both models are used for this analysis where the coefficients are $(a_2, a_1) = (0.15, 0.85)$. As can be seen, the capacity cost of the deterministic model is less than that of the robust model across all scenarios. The additional capacity cost in the robust model relative to the deterministic model is due to the demand uncertainty consideration in the robust model. The unmet/delayed penalty of the deterministic model is larger than the robust model across all the scenarios. In terms of overall cost C_T, the robust model outperforms the deterministic model, meaning that the consideration of additional capacity would increase the capacity cost but the added cost is offset by the delayed and unmet demand penalty costs. Figure 4.21 shows that across all scenarios, C_T in the robust model is less than C_T in the deterministic model.

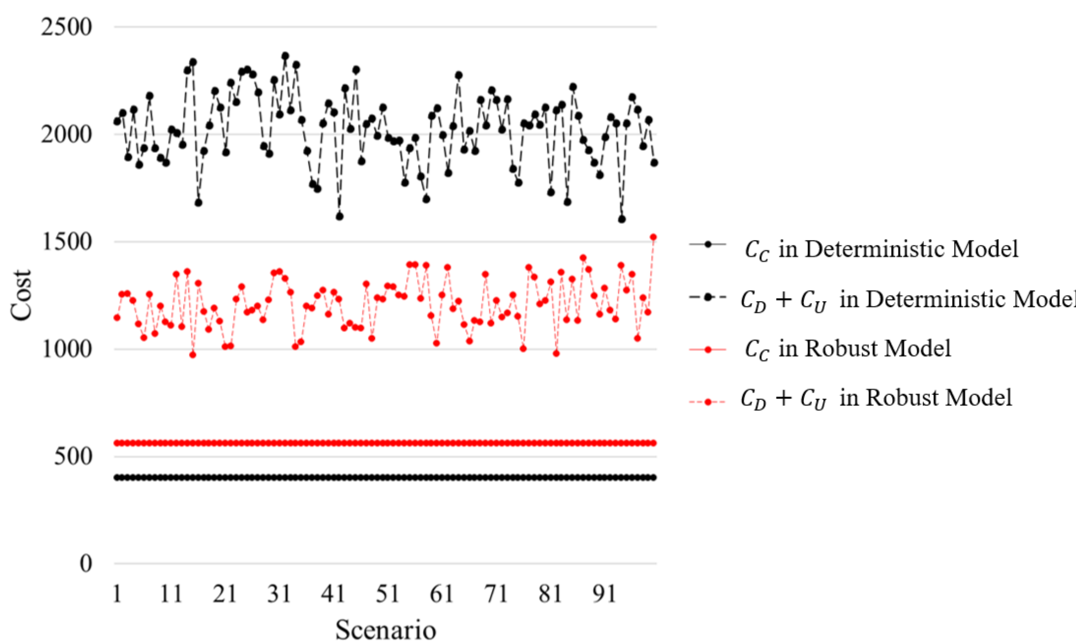


Figure 4.20: Cost Analysis for Robust vs. Deterministic Model

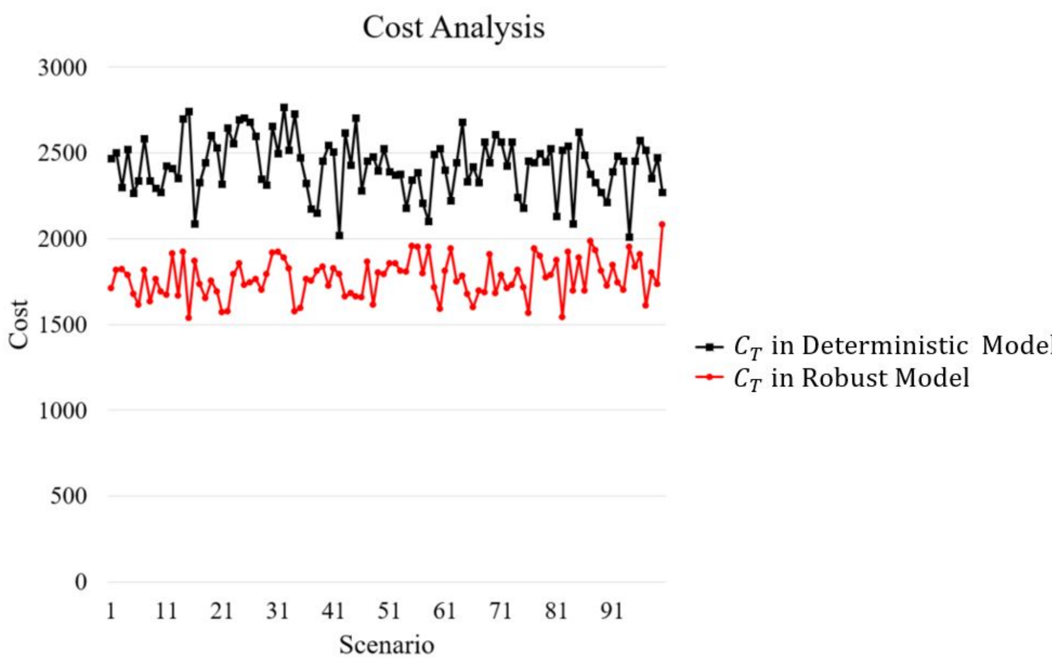


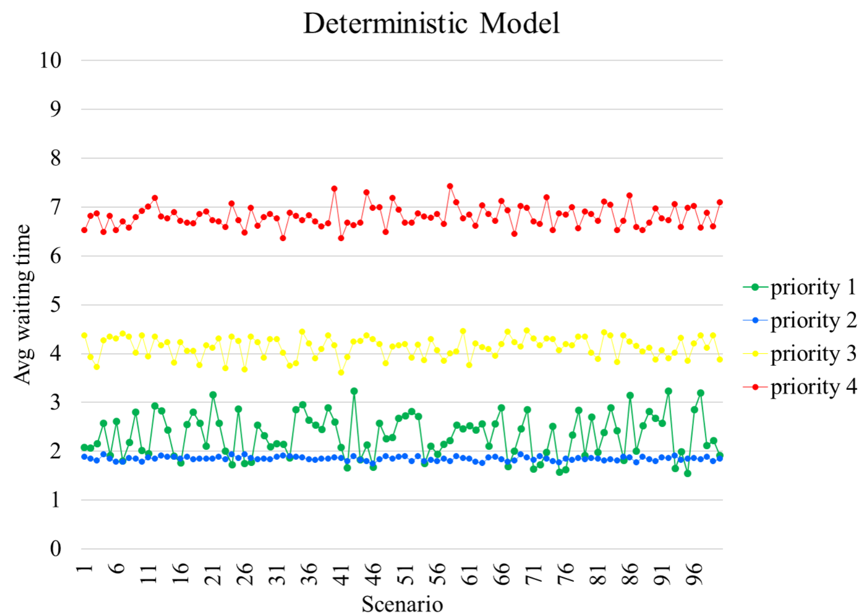
Figure 4.21: Total Cost Analysis for Robust vs. Deterministic Model

Figure 4.22a and 4.22b present another performance measure for comparing the two models, average wait time. Average wait time is defined for each priority level independently and is calculated as shown in equation (4.1), where $x_{t,n,p}$ is the number of patients with priority p who receive service after n days of waiting. It shows the average number of days that patients of priority p have been waiting until they receive service.

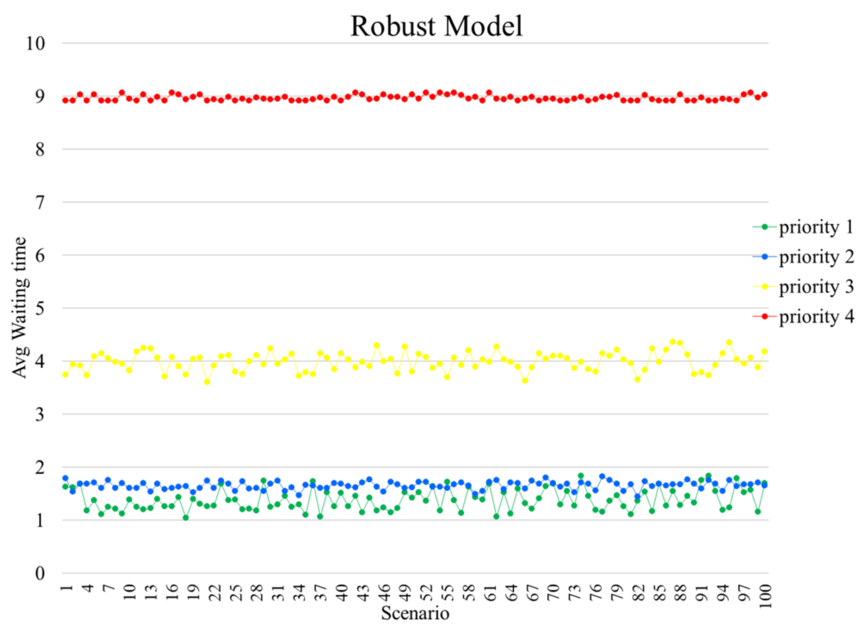
$$\text{Average wait time}_p = \frac{\sum_{t=1}^{T}\sum_{n=1}^{t}\left(x_{t,n,p} \times n\right)}{\sum_{t=1}^{T} x_{t,n,p}}, \quad \forall p \in \mathcal{P} \tag{4.1}$$

For each priority and each scenario, the average wait time throughout the horizon is presented. In the robust model, in general, average wait time increases with priority level, which means that priority 1 has the lowest average wait time and priority 4 has the highest average wait time of all. The same interpretation can also be made for the deterministic model except for a few exceptions for priority 1 and priority 2, where for some scenarios average wait time of priority 1 is slightly larger than that of priority 2. This inconsistency is rooted in the capacity distribution, the wait-time target, and the required service level differences. The demand for priority 1 is less than that of priority 2, which results in a higher capacity allocation to priority 2 relative to priority 1. The smaller capacity, combined with higher required service level and smaller wait-time target for priority 1 relative to priority 2, makes the recovery of a demand perturbation harder for priority 1 relative to priority 2. This is why priority 1 experienced a slightly higher average wait time under the deterministic model when the demand uncertainty is considered.

Across all scenarios and for the two most critical priorities, the average wait time of the robust model is less than that of the deterministic model. This means that the robust model favours critical priorities, which is more intuitive and suitable from a policy-making perspective. In general, the average wait time for priories 2, 3, and 4 across all scenarios are less than their associated wait-time targets; for priority 1, the average wait time is slightly more than the wait-time target due to the very small wait-time target for priority 1 – one day. Service level statistics are presented in Table 4.2.



(a) Avg wait time in Deterministic Model

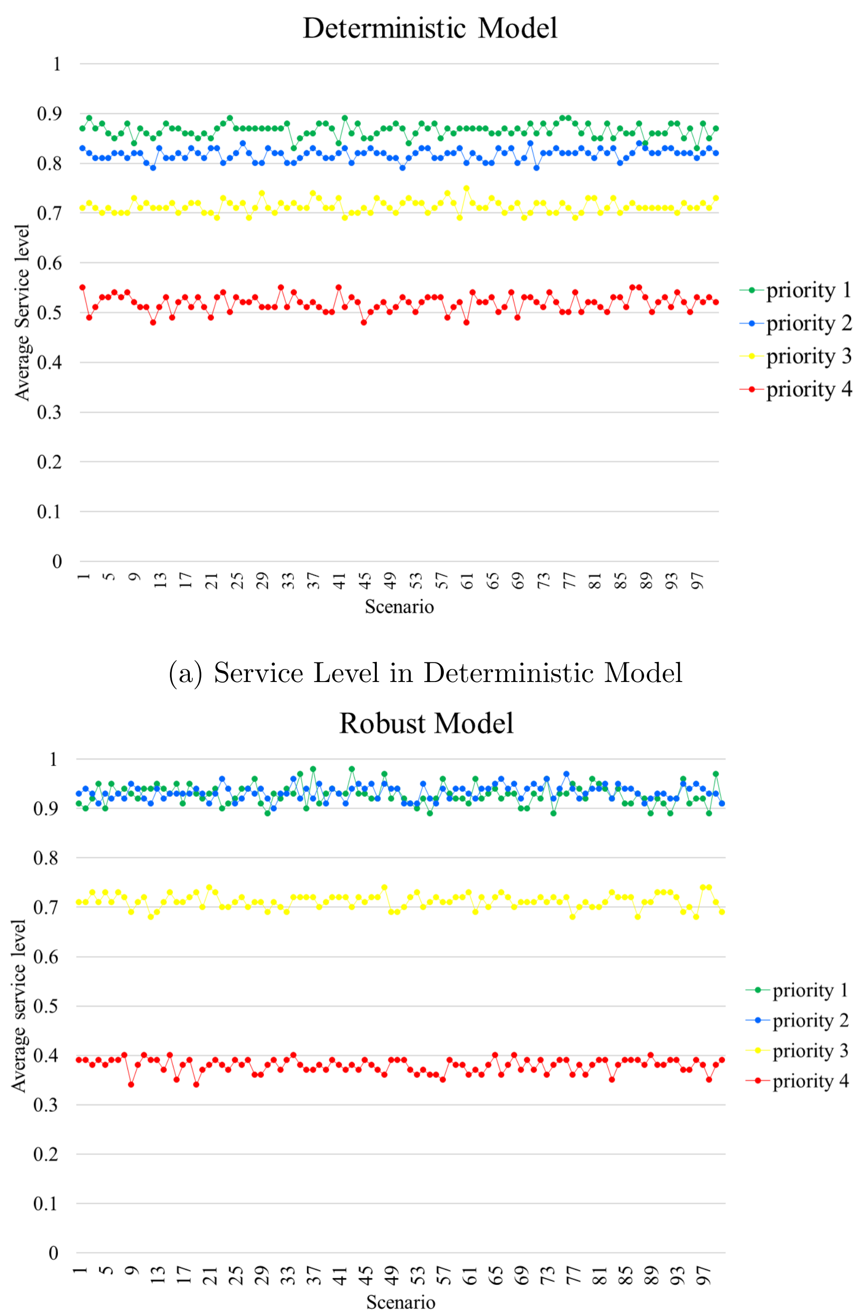


(b) Avg wait time in Robust Model

Figure 4.22: Avg wait time in Deterministic and Robust model

Service level for the two models and for all the tested scenarios are presented in Figure 4.23. Service level of each priority level is calculated using the equation (4.2):

$$\text{Service Level}_p(\%) = \Big(1 - \big(\frac{\sum_{t=1}^{T} \text{unmet demand}_{t,p}}{\sum_{t=1}^{T} \text{total demand}_{t,p}}\big)\Big) \times 100, \quad \forall p \in \mathcal{P} \tag{4.2}$$



(a) Service Level in Deterministic Model

(b) Service Level in Robust Model

Figure 4.23: Service Level in Deterministic and Robust Model

The statistics for the service levels of all the scenarios are presented in Table 4.2. Service level in the robust model for priority 1 and 2 is 93% and decreases for priority 3 and 4 to 71% and 38% respectively. In the deterministic model, the service level for priority 1 to 4 are 87%, 82%, 71%, and 52% respectively. Overall the robust model outperforms the deterministic model in terms of service time and across all priorities, except for priority 4. This exemption is yet another example that the robust approach favours critical patients with higher priority.

Table 4.2: Robust vs Deterministic, Service Level and wait time Analysis

Model	Statistical Measure	**Average wait time**				**Service Level**			
		P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4
Deterministic	Mean	2.33	1.85	4.13	6.81	87%	82%	71%	52%
	Std Dev	0.44	0.04	0.21	0.22	1%	1%	1%	2%
Robust	Mean	1.38	1.65	3.99	8.97	93%	93%	71%	38%
	Std Dev	0.2	0.08	0.17	0.05	2%	1%	1%	1%

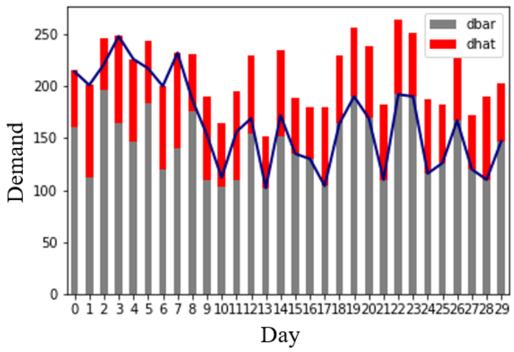
Table 4.3 represents the statistics of different cost components in the objective function. Although in the robust model the capacity cost, C_C considered in the model is higher, results confirm that by considering the demand uncertainty, the robust model outperforms the deterministic model since the penalty cost for unmet demand and delayed patients, $C_D + C_U$, is less in the robust model.

Table 4.3: Robust vs. Deterministic, Cost Analysis

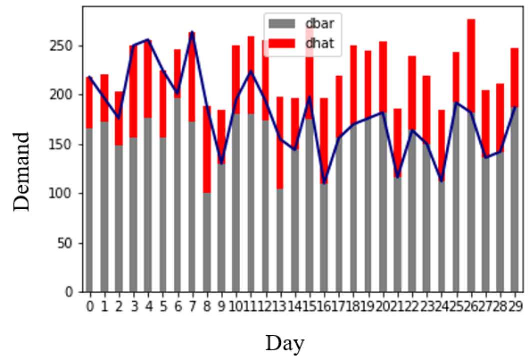
Model	Statistical Measure	Capacity Cost	Penalty Cost(cancelled/delayed)
Deterministic	Mean	401.85	2025
	Std Dev	0.00	166.63
Robust	Mean	562.5	1208.40
	Std Dev	0.00	113.4

4.3.2 Capacity Allocation Analysis

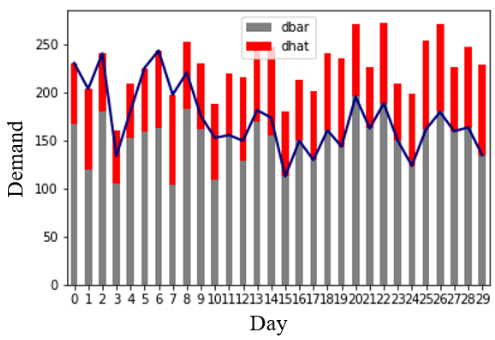
In this section, the sensitivity of the capacity allocations of the deterministic model and the robust model are analyzed and compared. Four random demands are generated based on the nominal demand, $\bar{d}$, and its possible deviation, $\hat{d}$, within the budget of uncertainty and the total capacity allocation of all priorities is evaluated across the planning horizon. Figure 4.24e illustrates the nominal demand used in this analysis. Figure 4.24a to 4.24d show the four random demands, D_1, D_2, D_3, and D_4 which are used in the robust model and deterministic model for this section analysis.



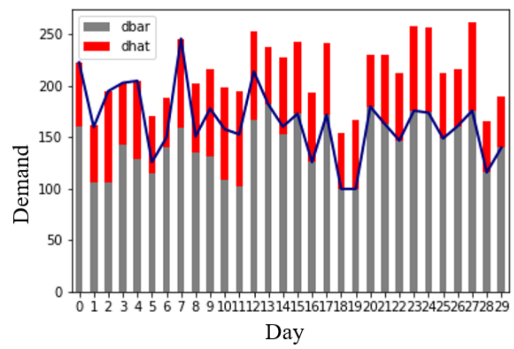
(a) Random Demand D_1



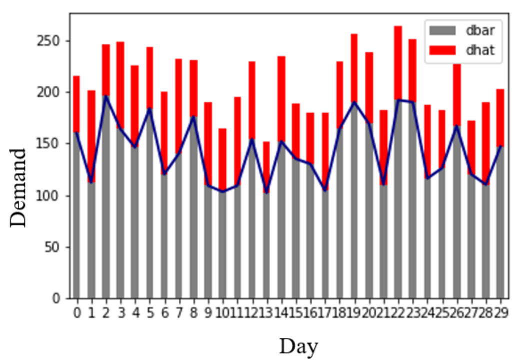
(b) Random Demand D_2



(c) Random Demand D_3



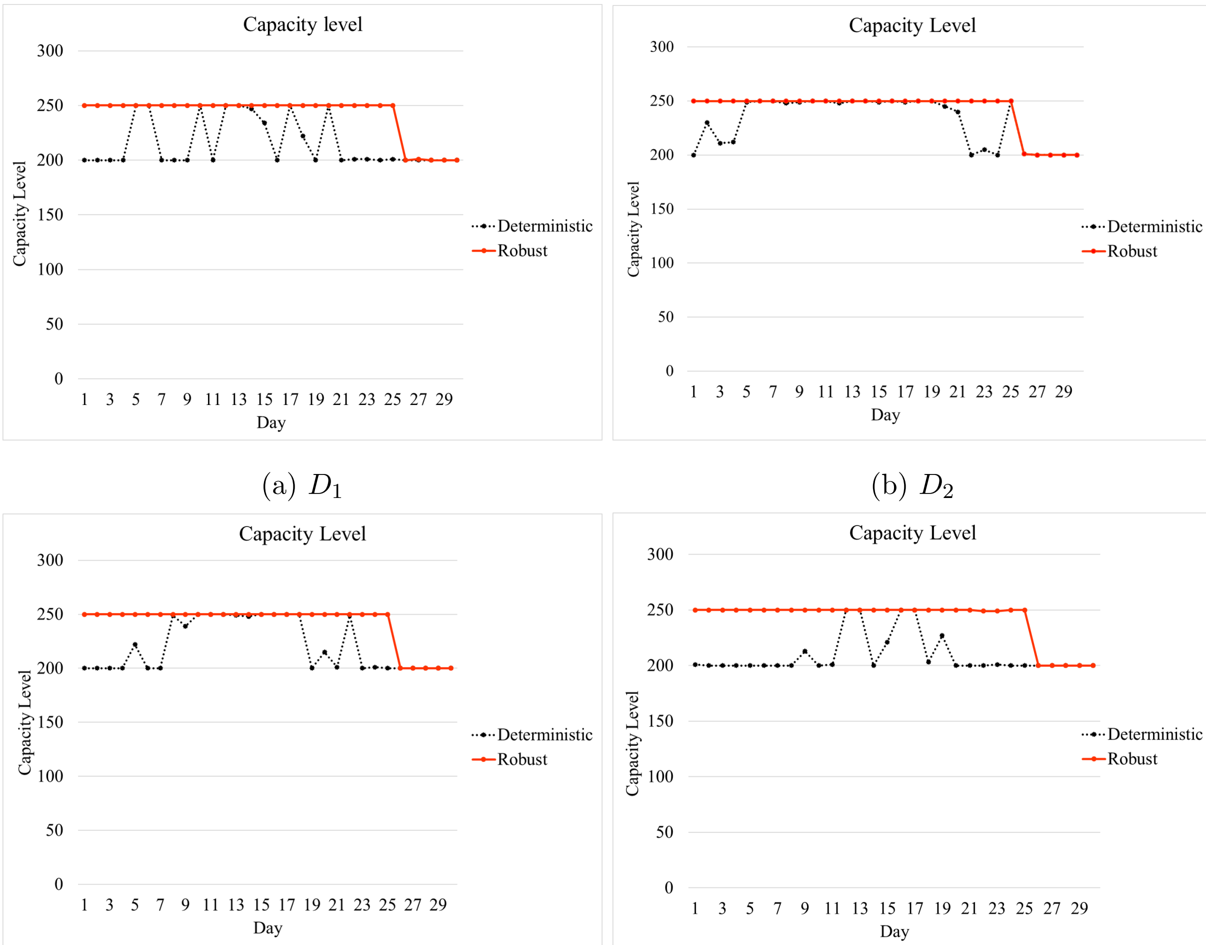
(d) Random Demand D_4



(e) Nominal Demand $\bar{d}$

Figure 4.24: Random Demands and Nominal Demand used for Capacity Allocation Analysis in Robust and Deterministic Models

The capacity allocations at the optimality of either model under different demands are extracted from the simulation models. The capacity allocations are plotted in Figure 4.25. Across all simulation results of different demands, (D_1, D_2, D_3, and D_4), the robust model has a higher capacity relative to the deterministic model, which is the result of considering demand uncertainty in the robust model. It is noted that the deterministic model's capacity allocation is more volatile when compared to that of the robust model. This higher volatility is the result of the deterministic model's attempt to respond to the specific demand. The less volatility (more stability) of the capacity allocation of the robust model can be associated with its consideration of the demand uncertainty throughout the planning horizon. From a policy planning and execution perspective, a less volatile capacity allocation is more favourable and easier to implement. Taking too extreme situations as an example, it is easier to implement and plan for a unique capacity level throughout the horizon rather than planning for different capacity levels for each day of the horizon to perfectly meet the demand. The volatile capacity is particularly harder to implement when internal working hours or space allocation arrangements are to be made to create additional capacity. Given that the demand uncertainty is inevitable and that capacity and delay in service costs can be measured in the field, for policy making and policy implementation, robust modelling could be a more suitable tool rather than deterministic modelling.



(a) D_1

(b) D_2

(c) D_3

(d) D_4

Figure 4.25: Capacity level obtained by Deterministic and Robust Model at different demands

Chapter 5

Conclusions and Future Studies

Inefficient healthcare infrastructure and policies have resulted in lengthy wait times for patients in Canada and worldwide. Aside from the economic losses that stem from the lengthy wait times, they cause risks to the lives of patients. Given the budget and resource limitations, creating new healthcare facility and infrastructure may not always be a feasible and sustainable solution to the healthcare inefficiencies; instead, the focus should be on optimizing the use of existing infrastructure through various policies and strategies to reduce inefficiencies in a healthcare system a cost-effective manner.

This research aims at healthcare policy development and quantitative analysis for practical applications. A multi-priority multi-period capacity allocation and patient scheduling problem is developed and solved using an adversarial-based algorithm.

A sensitivity analysis conducted on the modelling parameters leads to the conclusion that it is highly important to use calibration techniques or expert opinion to determine the modelling parameter values; it is shown in Section 4.2 that the results and policies extracted from an un-calibrated model are unreliable and misleading.

A comparative analysis of the modelling results leads to the conclusion that the robust model outperforms the deterministic model in terms of the overall cost, which includes unmet demand penalty cost, delayed patients penalty cost, and capacity cost. Although the robust model may result in a more conservative capacity and hence is more costly, the cost savings from delayed patients penalty costs and unmet demand penalty costs offset the effect of the capacity costs.

Compared to the deterministic approach, the robust approach leads to a more stable and less fluctuant capacity allocation through the planning horizon, which makes the robust policy easier to execute and administer; the robust modelling is hence considered a more suitable tool for policy planning purposes. The unpredictability and inevitability of demand uncertainties along with the estimability of capacity and patient delay costs naturally favours robust modelling over deterministic modelling for policy analysis and planning purposes.

This study sheds light on a few avenues for future research. First, the proposed robust modelling approach and the solution methodology can be used for evaluating adaptive scheduling policies where the scheduling schemes are updated throughout the horizon based upon feedback from previous scheduling. Second, using the proposed modelling approach, one can examine the performance of scheduling policies and compare the results against benchmark scheduling policies (e.g., first come first served, etc.). Lastly, upon the availability of data, empirically calibrated capacity cost functions along with estimated demands and their deviations can be used as input to the proposed model to drive scheduling policies and insights for practical purposes.

References

Bernardetta Addis, Giuliana Carello, and Elena Tànfani. A robust optimization approach for the operating room planning problem with uncertain surgery duration. In *Proceedings of the international conference on health care systems engineering*, pages 175–189. Springer, 2014a.

Bernardetta Addis, Giuliana Carello, and Elena Tànfani. A robust optimization approach for the advanced scheduling problem with uncertain surgery duration in operating room planning-an extended analysis. 2014b.

Amir Ahmadi-Javid, Zahra Jalali, and Kenneth J Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3–34, 2017.

Nazanin Aslani, Onur Kuzgunkaya, Navneet Vidyarthi, and Daria Terekhov. A robust optimization model for tactical capacity planning in an outpatient setting. *Health Care Management Science*, pages 1–15, 2020.

Davood Astaraky and Jonathan Patrick. A simulation based approximate dynamic programming approach to multi-class, multi-resource surgical scheduling. *European Journal of Operational Research*, 245(1):309–319, 2015.

Norman TJ Bailey. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society: Series B (Methodological)*, 14(2):185–199, 1952.

Aharon Ben-Tal and Arkadi Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical programming*, 88(3):411–424, 2000.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton University Press, 2009.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Operations research*, 52(1): 35–53, 2004.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Hans-Georg Beyer and Bernhard Sendhoff. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

Daniel Bienstock. Histogram models for robust portfolio optimization. *Journal of computational finance*, 11(1):1, 2007.

Daniel Bienstock and Nuri ÖZbay. Computing robust basestock levels. *Discrete Optimization*, 5(2):389–414, 2008.

Brecht Cardoen, Erik Demeulemeester, and Jeroen Beliën. Operating room planning and scheduling: A literature review. *European journal of operational research*, 201(3):921–932, 2010.

Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: a review of literature. *Production and operations management*, 12(4):519–549, 2003.

Santanu Chakraborty, Kumar Muthuraman, and Mark Lawley. Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions*, 42 (5):354–366, 2010.

CIHI. Canadian Institute for Health Information. http://waittimes.cihi.ca/procedure/mri?show=5090#year, 2016. Accessed: 2020-10-27.

CIHI. Canadian Institute for Health Information. http://waittimes.cihi.ca/procedure/mri?show=5090#year, 2017. Accessed: 2020-10-27.

CIHI. Canadian Institute for Health Information. http://waittimes.cihi.ca/All#year, 2019. Accessed: 2020-10-27.

George B Dantzig. Linear programming under uncertainty. *Management science*, 1(3-4): 197–206, 1955.

Brian Denton and Diwakar Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE transactions*, 35(11):1003–1016, 2003.

Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on matrix analysis and applications*, 18(4):1035–1064, 1997.

Fraser Institute. Waiting Your Turn: Waiting Times for Health Care in Canada. https://www.fraserinstitute.org/studies/waiting-your-turn-wait-times-for-health-care-in-canada-2019, 2019. Accessed: 2020-10-27.

Virginie Gabrel, Cécile Murat, and Aurélie Thiele. Recent advances in robust optimization: An overview. *European journal of operational research*, 235(3):471–483, 2014.

Yigal Gerchak, Diwakar Gupta, and Mordechai Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3): 321–334, 1996.

Yasin Gocgun and Archis Ghate. Lagrangian relaxation and constraint generation for allocation and advanced scheduling. *Computers & Operations Research*, 39(10):2323–2336, 2012.

Yasin Gocgun and Martin L Puterman. Dynamic scheduling with due dates and time windows: an application to chemotherapy patient appointment booking. *Health care management science*, 17(1):60–76, 2014.

Matias Holte and Carlo Mannino. The implementor/adversary algorithm for the cyclic and robust scheduling problem in health-care. *European Journal of Operational Research*, 226(3):551–559, 2013.

Yangzi Jiang, Hossein Abouee-Mehrizi, and Yuhe Diao. Data-driven analytics to support scheduling of multi-priority multi-class patients with wait time targets. *European Journal of Operational Research*, 281(3):597–611, 2020.

Kenneth J Klassen and Thomas R Rohleder. Scheduling outpatient appointments in a dynamic environment. *Journal of operations Management*, 14(2):83–101, 1996.

Kenneth J Klassen and Thomas R Rohleder. Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. *International Journal of Service Industry Management*, 2004.

Mehdi Lamiri, Xiaolan Xie, Alexandre Dolgui, and Frédéric Grimaud. A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research*, 185(3):1026–1037, 2008.

David V Lindley. The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 48, pages 277–289. Cambridge University Press, 1952.

Nan Liu, Serhan Ziya, and Vidyadhar G Kulkarni. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2):347–364, 2010.

James M Magerlein and James B Martin. Surgical demand scheduling: A review. *Health services research*, 13(4):418, 1978.

Houra Mahmoudzadeh, Akram Mirahmadi Shalamzari, and Hossein Abouee Mehrizi. Robust multi-class multi-period patient scheduling with wait time targets. *Operations Research for Health Care*, page 100254, 2020.

Ho-Yin Mak, Ying Rong, and Jiawei Zhang. Appointment scheduling with limited distributional information. *Management Science*, 61(2):316–334, 2015.

Kumar Muthuraman and Mark Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *Iie Transactions*, 40(9):820–837, 2008.

Mahshid Salemi Parizi and Archis Ghate. Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Computers & Operations Research*, 67: 90–101, 2016.

Jonathan Patrick, Martin L Puterman, and Maurice Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 56(6):1507–1525, 2008.

Zbigniew H Przasnyski. Operating room scheduling: A literature review. *AORN journal*, 44(1):67–82, 1986.

Antoine Saure, Jonathan Patrick, Scott Tyldesley, and Martin L Puterman. Dynamic multi-appointment patient scheduling for radiation therapy. *European Journal of Operational Research*, 223(2):573–584, 2012.

Antoine Sauré, Mehmet A Begen, and Jonathan Patrick. Dynamic multi-priority, multi-class patient scheduling with stochastic service times. *European Journal of Operational Research*, 280(1):254–265, 2020.

Allen L Soyster. Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations research*, 21(5):1154–1157, 1973.

Van-Anh Truong. Optimal advance scheduling. *Management Science*, 61(7):1584–1597, 2015.

Richard M Van Slyke and Roger Wets. L-shaped linear programs with applications to optimal control and stochastic programming. *SIAM Journal on Applied Mathematics*, 17(4):638–663, 1969.