Applying an Integer Linear Programming Model to an appointment scheduling problem

Dissertação de Mestrado

Eduardo Silva Medeiros

Mestrado em

Ciências Económicas e Empresariais

Ponta Delgada
2021

Applying an Integer Linear Programming Model to an appointment scheduling problem

Dissertação de Mestrado

Eduardo Silva Medeiros

Orientadores

Prof. Doutor Francisco José Ferreira Silva
Prof. Doutor Pedro Portugal de Sousa Nunes

Dissertação submetida como requisito parcial para obtenção do grau de Mestre em Ciências Económicas e Empresariais, com especialização em Economia e Políticas Públicas.

ABSTRACT

Outpatient appointment management can be a complex process since it involves many conflicting stakeholders. As for the patients it might be important to minimize waiting time. Simultaneously, for healthcare workers, fair working conditions must be guaranteed. Thus, it is increasingly necessary to have workload balance and resource optimization as the main concerns in the scheduling and planning of outpatient appointments.

In this dissertation, a two-model approach for designing an appointment scheduling is proposed. This approach is formulated as two mathematical Integer Linear Programming models that integrate the objective of minimizing workload difference and improving workload balance.

The models were structured and parameterized according to randomly generated data. For this, the work was developed in Java, generating said data. Model I minimizes the workload differences among rooms. Model II, on the other hand, proposes a new objective function that minimizes the maximum workload difference, with a $minxmax$ decision process.

The computational models behaves efficiently in reasonable run times for numerical examples with less than approximately 10 rooms available. Higher run times are observed when numerical examples surpass these number of available rooms. Regarding workload balance, it was observed that the number of specialties available for appointments and the demand for each day were the most influential in the minimization of workload difference.

Model II results show a shorter model run time and more optimal solutions. As the differences between both Models are not considerable, Model I might propose a better set of solution for decision makers since it minimizes the total workload difference amongst rooms instead of only minimizing the maximum workload difference between any two rooms.

Keywords: Operational research, Healthcare, Appointment scheduling, Outpatient clinics, Integer Linear Programming

RESUMO

A gestão de consultas ambulatórias pode ser um processo complexo, uma vez que envolve vários stakeholders com diferentes objetivos. Para os utentes poderá ser importante minimizar os tempos de espera. Simultaneamente, para os trabalhadores do setor da saúde, condições de trabalho justas devem ser garantidas. Assim, é cada vez mais necessário ter em conta o equilíbrio de cargas horárias e a otimização dos recursos disponíveis como principais preocupações no agendamento e planeamento de consultas.

Nesta dissertação, uma abordagem com dois modelos para a criação de um sistema de agendamento de consultas é proposta. Esta abordagem é feita em programação linear, com dois modelos que têm como objetivo minimizar as diferenças de cargas horárias e melhorar o seu equilíbrio ao longo do planeamento.

Os modelos foram estruturados e parametrizados de acordo com dados gerados aleatoriamente. Para isso, o desenvolvimento foi feito em Java, gerando assim os dados referidos. O Modelo I minimiza as diferenças de carga horária entre os quartos disponíveis. O Modelo II, por outro lado, propõe uma nova função objetivo que minimiza a diferença máxima observada, com um processo de decisão $minxmax$.

Os modelos mostram resultados eficientes em tempos de execução razoáveis para instâncias com menos de aproximadamente 10 quartos disponíveis. Os tempos de execução mais altos são observados quando as instâncias ultrapassam este número de quartos disponíveis. Em relação ao equilíbrio da carga horária, observou-se que o número de especialidades disponíveis para atendimento e a procura por dia foram o que mais influenciou a minimização da diferença da carga horária.

Os resultados do Modelo II mostram melhor tempo de execução e um maior número de soluções ótimas. Uma vez que as diferenças entre os dois modelos não são consideráveis, o Modelo I poderá representar um melhor conjunto de soluções para os decisores já que minimiza a diferença da carga horária total entre quartos em vez de apenas minimizar o valor máximo da diferença de carga horária entre quaisquer dois quartos.

Palavras-chave: Investigação operacional, Saúde, Agendamento de consultas, Clínicas, Programação linear

To all those who inspired it and will not read it.

ACKNOWLEDGEMENTS

I would like to express my gratitude and appreciation for my supervisors Prof. Doutor Francisco José Ferreira Silva and Prof. Doutor Pedro Portugal de Sousa Nunes, whose guidance, support and expertise in the field were invaluable throughout this study.

My appreciation also goes out to the Faculdade de Economia e Gestão for all the considerate contributions.

A special word of thanks goes out to my mother. I cannot express enough how grateful I am for you and everything you have done for me and my future.

To Mariana: Thank you for always walking besides me, and above all, for never letting me give up on anything, however difficult it may be. I could never guess I would be so lucky to have someone like you in my life. We go together, you and me.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I - INTRODUCTION

Healthcare is an important right to everyone and it must be provided quick and efficiently. Accordingly to Ahmadi Javid, Jalali, and Klassen (2016), developing efficient healthcare systems has become more important in the last few decades for two major reasons: 1) the rapid increase in healthcare expenditures in more developed countries, and 2) the simultaneous growth of demand for healthcare services and patients' expectations of service quality (Hulshof, Kortbeek, Boucherie, Hans, & Bakker, 2012).

Indeed, patients continue to demand a good quality of life especially by demanding shorter waiting times. In the private health care sector, this means that competition between service providers is escalating and in the public sector, workloads become congested and inefficient work methods are more often common than not. An imbalance in the workplace can lead to over-productivity by some team members and under-productivity by others. A lack of workload balance can also lead to employee dissatisfaction and an overall decrease in organization morale. In addition to that, public sector co-workers are migrating towards the private sector, based on these unfair working conditions, low wages and increased overtimes (Publico, 2017).

This results in a challenging situation for outpatient service providers (i.e. in contrast to "inpatients" who are admitted and stay in the hospital). Although some types of services involve many aspects of quality other than waiting time, patients use waiting time, good professionals and adequate equipment, and an overall good customer experience as a differentiating factor among their choice for service providers.

As a result, the study of efficient health care systems is becoming more popular and the field for such research and development is operational research, which provides numerous methodologies and solution techniques. Within the particular case of outpatient clinics, appointment systems (AS) are the most important components for efficient care delivery. Outpatient appointment system (OAS) problems are an attractive research area, having been studied for more than half a century, starting with the pioneer study of Bailey (1952).

There are a lot of classifications for Outpatient Appointment System decisions in the literature. Appointment rules, patient classification and characterization, environmental factors and many others. As shown in Figure 1, when addressing the problem of Outpatient Appointment Systems, these level of decisions can be: strategic, tactical or operational. Strategic decisions are long-term decisions that determine the base structure of an Outpatient Appointment System. Tactical decisions are medium-term decisions related to how patients as a whole are scheduled, or how groups of patients are processed. Operational decisions are those decisions that are adjusted more frequently in correspondence to the current external and internal conditions. All these decisions allow constructing a complete schedule for the appointment system design in clinics.

Bailey (1952) found that most medical services do not schedule any more than 20 or 30 clients in a room session; this is still true today. He also determined that the best scheduling policy is to place two patients in the first appointment slot and spread the rest evenly over the period based on average service times. Ho and Lau (1992) identified eight "best" rules by placing them on an efficient frontier such that each rule dominated all other rules in terms of server idle time, client waiting time, or some combination of both. Based on this, the appropriate rule can be chosen by a service provider depending on the relative weight they give to server idle time versus client waiting time. Other studies also extended their prior theoretical work by considering environmental conditions such as service time distribution, number of patients per session, no-show probability, walk-ins and many others.

Generally, waiting lists can be managed in many ways. In most cases, the guidelines to manage waiting lists are according to patients' priority and waiting time. In practice, when a patient requests an appointment, a scheduler in the clinic first compares the current schedule of each time set (daily, weekly, monthly...) to the scheduling template to find available appointment time slots, and then schedules an appointment at one of the available times. The most common issues with this is that there are patient no-shows and large variation of service times. For example, nearly half of the scheduled appointments are missed for certain service type and the service times vary from a few minutes to more than an hour (Qu, Peng, Kong, & Shi, 2013). The service time is defined as the actual time a patient spends with a healthcare provider in the consultation and examination.

There are many appointment systems and scheduling policies. Dynamic demands need dynamic appointment systems and an efficient workload is the starting point to improve and aim at optimizing all other problems. One of the major goals in outpatient appointment scheduling is to reduce several time-based system performance measures including provider idle time, overtime, and patient waiting time, and thus, improve efficiency.

To generate appropriate and optimized schedules, it is necessary to gather all relevant variables to the problem and to construct a mathematical model that is a sufficiently precise representation of the situation that need solving. The model should focus on the hospital characteristics to predict the best way to allocate resources, which on this case were, specialties and patients.

The main barrier when modeling is to define a trade-off between the robustness of the model and an acceptable complexity. It is also necessary to consider the visions of the various stakeholders: patients, medical staff and administration. The different stakeholders are driven by different interests: for example, the main interest for patients will be to reduce the waiting time; for the hospital staff the greatest concern is related to the fairness of workload distributions; while administration keeps the focus of managing costs and all the financial issues.

This dissertation focuses on a general modeling approach with the objective of balancing these problems within medical outpatient appointment scheduling, providing fair work conditions for the healthcare professionals and while still having efficient systems that improve the patients' customer experience.

Some existing problems in clinics, such as the ones previously referred, are common issues in many outpatient clinics. Moreover, more restrictions on outpatient appointment scheduling should be considered due to the diversity of services provided and resources available within a work day.

This dissertation will not only address the common issues on outpatient appointment scheduling, but also consider extraordinary factors that combined together, improve efficiency and workflows. The aforementioned issues and characteristics motivated the focus on multi-category outpatient appointment scheduling problems, for which a two-model approach of minimizing workload differences is proposed.

Model I has the objective of minimizing the total workload differences among rooms.

On the other hand, Model II proposes a new objective function that minimizes the maximum workload difference, with a $minxmax$ decision process.

Both in Model I and Model II, each room session is assigned to one of the given specialties and specified with the number of appointments for each service type belonging to the assigned specialty.

The utility of these models is to support the making of decision support models within healthcare environments, being either clinics or hospitals. It allows healthcare managers to take informed decisions on the number of appointments for certain clinical specialties and service types that should be planned for any given session, as well as the total workload for available rooms and resources, in order to minimize workload differences and guarantee fairness among healthcare professionals.

This dissertation is organized as follows. Chapter II develops a literature review, which reports on the approaches proposed by several authors for outpatient appointment scheduling. There are many studies related to the topic under study, but a selection of some papers is made with the aim of showing that the different chosen approaches may be consequence of the numbers of factors to consider, hence, having the necessity of a general and broad model. Chapter III includes the suggested methodology approached towards the problem defined and the description of the Integer Linear Programming models proposed to cover the appointment scheduling problem, with the two approaches previously mentioned and their respective formulation. Chapter IV presents the numerical example generation process used for the computational experiments. The data is used to validate the models and perform sensitivity analysis evaluating how some changes in the data or in the mathematical formulation impact the results, as well as a comparison between the outputs of Model I and Model II, as described in Chapter V. The main results

are presented and analyzed in this chapter with an extended discussion on all the decisions taken and their impact on the results. These results intend to answer the research question of this dissertation on how to minimize workload differences, and what is the best approach to do it. Finally, Chapter VI concludes this dissertation, highlights the limitations of the project and outlines suggestions for future work.

CHAPTER II - RELATED LITERATURE

The Operational Research (OR) literature on appointment scheduling is fairly extensive since Bailey (1952) first introduced it. According to this author, effective scheduling systems aim to match demand with the capacity of health systems, so that resources are better used and patients' waiting times are minimized.

The literature for Outpatient Appointment Scheduling has been analyzed in several comprehensive literature reviews such as Cayirli and Veral (2003); Gupta and Denton (2008) and Ahmadi Javid et al. (2016). The main goal of the literature review undertaken by Cayirli and Veral (2003) is to review prior formulations and modeling considerations for Outpatient Appointment Systems, whereas the one by Gupta and Denton (2008) focus on describing the most common types of healthcare appointment systems, paying particular attention to the factors complicating Outpatient Appointment Systems planning. Most recently, Ahmadi Javid et al. (2016) provide a comprehensive review of recent analytical and numerical optimization studies.

A myriad of classifications for Outpatient Appointment Systems decisions has been proposed in the literature. Cayirli and Veral (2003), for instance, view Appointment Systems design as a series of three decision levels, namely "appointment rule," "patient classification," and "adjustments for no-shows and walk-ins.", and state that at the minimum, the decision is reduced to finding the best appointment rule, which is the basic template that specifies the combination of block sizes and appointment interval lengths. Hulshof et al. (2012) present a classification to review the operational research and management science (OR/MS) studies related to planning decisions in healthcare. For ambulatory care services, they list seven key decisions: number of patients per room, patient overbooking, length of the appointment interval, number of patients per appointment slot (i.e., block size), sequence of appointments, queue discipline in the waiting room, and anticipation for unscheduled patients. W. Y. Wang and Gupta (2011) divide Outpatient Appointment Systems decisions into two categories: clinic profile setup and appointment booking, based on a two-stage process that is usually used in outpatient clinics. Ahmadi Javid et al. (2016) adopt all these categorizations in their Outpatient Appointment Systems classification, introducing a broader framework which is organized according to whether the decisions made in designing and planning them are strategic, tactical, or operational.

Based on the structure proposed by Ahmadi Javid et al. (2016), this literature review merges the classifications made and is structured as follows:

2.1) Appointment System Design: Considers decisions made to design and plan Outpatient Appointment Systems, but focusing only on medium and long-term planning, namely strategic and tactical levels of decisions;

2.2) Environmental Factors: Breaks down the factors complicating Outpatient Ap-

pointment Systems planning;

2.3) Solution Methods and Modeling Approaches: States the most common solution methods and modeling approaches.

2.1 Appointment System Design

The Appointment Systems is usually applied when admitting patients to hospitals or outpatient clinics. Cayirli and Veral (2003) state that most literature on appointment scheduling can be split into two case scenarios. On the one hand, the ones whose appointments are made prior to the beginning of a clinic session (static). On the other hand, schedules where future arrivals are revised continuously over the course of the day based on the current state of the systems (dynamic).

Gupta and Denton (2008) classify appointment scheduling into two types on the basis of the type of waiting modelled, particularly, direct or indirect. They state that indirect waiting time is the difference between the time that a patient requests an appointment and the time of said appointment and that direct waiting time is the difference between a patient's appointment time (or his/her arrival time if he/she is late) and the time when he/she is actually served by the service provider.

In healthcare, a significant emphasis has been put on increasing patient satisfaction and minimizing access and waiting times. The overall problem is to control waiting times by applying an appropriate, sensitive, and responsive appointment procedure in which the estimated patient flow corresponds to the available resources. However, not all literature dedicate their work to patient-centered operations in hospitals (Marynissen & Demeulemeester, 2019). Indeed, because hospitals need to become more cost-efficient and face budget cuts, profit maximization is becoming more popular. It is possible to maximize profit by maximizing the number of patients scheduled, maximizing the contribution margin or minimizing the idle time of resources. Another approach is considering fairness measures in the Appointment Systems design. Although only few studies contemplate it in their objective functions, balancing workloads, taking congestion into account and/or doctor/patient fairness procedures can make a notable difference, indirectly, in numerous factors such as reducing waiting and idle times, improving patient and doctor satisfaction, and even making hospitals more cost-efficient with a better organized methodology.

All these objective function types can be valid, depending on the context. The following subsections will discuss Outpatient Appointment Systems design and decision making. It is clear that all decision levels are important when focusing on Appointment Systems. However, not all decision levels are equally relevant for the purposes of this dissertation. Since this work will focus on medium and long-term planning, only strategic and tactical levels of decisions are displayed. Operational decisions are outside the scope

of this study, as they represent short-term concerns with efficiently scheduling individual patients. They should be variable, situational and dependent of the procedures adopted by the clinic.

2.1.1 Strategic Decisions

Strategic-level options for access policy, namely strategic (or design) decisions, will be discussed in this section. These are long-term decisions that determine the base structure of an Outpatient Appointment System.

i. Access policies for scheduled patients

In terms of scheduling process, there are three main types of scheduling policies: traditional, open-access and hybrid (L. W. Robinson & Chen, 2010). This means that scheduled patients, who make an appointment before arriving at the clinic, can be divided into two sub-classes: pre-scheduled patients, who are scheduled in advance of their appointment days (traditional) and same-day patients, who are scheduled on the same day that they call for an appointment (open-access). A policy that allows both is considered hybrid (i.e., accommodating both same-day and pre-scheduled appointments). In Cayirli and Veral (2003) review, the authors assume that within the pre-scheduled sub-class, the patients are most commonly served as first-come, first-served (FCFS).

According to these policies, the analyzed papers were classified in Table 1.

Table 1: Overview of different types of access policies for scheduled patients

Access Policy	References
Traditional	(Zacharias & Pinedo, 2014); (Turkcan, Zeng, & Lawley, 2012); (L. Robinson & Chen, 2003); (Kuiper, Kemper, & Mandjes, 2015); (K. J. Klassen & Yoogalingam, 2014); (K. Klassen & Yoogalingam, 2013); (K. J. Klassen & Yoogalingam, 2009); (Kaandorp & Koole, 2007); (Erdogan & Denton, 2013); (Denton & Gupta, 2003)
Open-access	(Muthuraman & Lawley, 2008); (Hulshof et al., 2012); (Erdogan & Denton, 2013)
Hybrid	(W. Y. Wang & Gupta, 2011); (Gupta & Denton, 2008); (Cayirli & Gunes, 2014); (Balasubramanian, Biehl, Dai, & Muriel, 2013)

Traditional policy often involve higher no-show rates, mainly because of longer indirect waiting times (Qu, Rardin, Williams, & Willis, 2007). An open-access policy can be proposed to avoid the negative effects of high no-show rates under a traditional policy (Murray & Tantau, 2000). However, daily fluctuations in patient demand – which may lead to poor resource utilization – and patient appointment-booking preferences are

barriers to the extensive use of open-access policies. Gupta and Wang (2008) show that open-access systems perform worse when there is greater variability of same-day demand or greater positive dependence among the same-day demands for different doctors. Different demand patterns and reserved slots for urgent patients are also described by K. J. Klassen and Rohleder (2004). L. W. Robinson and Chen (2010) compare the performance of open-access and traditional policies when the number of appointments are given, appointment time is deterministic, and patients are punctual. Their numerical analysis reveals that the open-access policy significantly out-performs the traditional policy in most cases.

ii. Walk-ins

Walk-in patients - commonly referred as walk-ins - are patients who arrive at the clinic without an appointment during the consultation session. Two major walk-in patient classes are considered in the literature: urgent and regular (Cayirli & Veral, 2003). The urgent walk-in patients often need to be treated as soon as possible, whereas regular walk-ins have a lower priority in the system; general approach is to leave slots open for potential walk-ins or wait for no-show slots. It should be noted that walk-in patients are different from same-day (or open-access) patients (L. W. Robinson & Chen, 2010).

Table 2: Overview of walk-in acceptance policy

Walk-in Acceptance	References
Yes	(Qu, Peng, Shi, & LaGanga, 2015), (Koeleman & Koole, 2012), (Cayirli & Gunes, 2014)
No	(Zacharias & Pinedo, 2014), (W. Y. Wang & Gupta, 2011), (Turkcan et al., 2012), (L. Robinson & Chen, 2003), (Kuiper et al., 2015), (K. J. Klassen & Yoogalingam, 2014), (K. Klassen & Yoogalingam, 2013), (K. J. Klassen & Yoogalingam, 2009), (Kaandorp & Koole, 2007), (Hulshof et al., 2012), (Gupta & Denton, 2008), (Erdogan & Denton, 2013), (Denton & Gupta, 2003), (Balasubramanian et al., 2013)

Koeleman and Koole (2012) study the appointment scheduling problem considering urgent walk-ins that arrive following a non-stationary Poisson process. Their analysis show that it is the best to leave slots open for urgent patients toward the end of the consultation session. K. J. Klassen and Rohleder (2004) also consider urgent walk-ins. The specific slots chosen by them are in the middle of the morning and the middle of the afternoon. They conclude that trade-offs exist with the choice: if more urgent slots are left earlier in the session, average patient waiting time is lower but fewer urgent patients are served; whereas if more slots are left later, doctor idle time is lower and more urgent patients are served.

Excluding these considerations, presence of walk-ins (either emergency and regular) is usually neglected in Outpatient Appointment Systems studies. The presence of walk-in patients leads to increased modeling complexity because of the dynamic stochastic arrivals of walk-in patients (Ahmadi Javid et al., 2016). This may be the reason why a majority of authors do not include them in their models.

Although walk-ins are an important factor, clinics are primarily used for consultation services for patients referred to them by the general doctors', and walk-ins are rarely accepted within their policies. This means that walk-ins must be anticipated and planned if clinics intend to consider them.

iii. Number of servers/resources

When entering in an hospital to go to an appointment, a patient usually follows a certain path. During this path they may deal with a variable number of resources - in addition to just seeing the doctor - such as clinic staff (i.e., nurses, and other medical specialists), medical equipment (e.g., medical imaging machines) (Ahmadi Javid et al., 2016; Marynissen & Demeulemeester, 2019). Hence, each resource type can either consist of a single server or multiple servers (Van de Vrugt, Noelle Maria, 2016).

In Cayirli and Veral (2003) review, the authors state that almost all studies in Outpatient Appointment Systems model a single-stage system where patients queue for a single service. However, in recent years, an increasing number of researchers started to acknowledge the multiple diagnostic problems when focusing on patient scheduling. The result is a series of research efforts to study the scheduling process on multiple resources.

In reality, patients must often undergo multiple diagnostic tests, consultations and/or surgeries to be treated. Therefore it is only logical that the number of studies in this field of study increases, which can be observed in Table 3.

iv. Types of Scheduling

Two scheduling approaches can be considered in Outpatient Appointment Systems: online (i.e., sequential) and offline (i.e., simultaneous). In the offline approach, appointments are scheduled after all requests have arrived, while in the online approach, patients are scheduled immediately upon the arrival of their request (Zacharias & Pinedo, 2014).

Indeed, upon receiving a request for service, schedulers have two options regarding their response time to the request. On the one hand, they can respond immediately with a date and time for the requested appointments. This implies that scheduling becomes a sequential process in which patients are given appointments in the order of the arrival time of their request. On the other hand, schedulers might also want to wait and collect requests for appointments in a waiting list, after which an algorithm is applied to select patients from this list (Marynissen & Demeulemeester, 2019).

Online systems are more common in practice, while the offline approach has received

Table 3: Overview of server type

Server type	References
Single	(Bailey, 1952), (Zacharias & Pinedo, 2014), (L. Robinson & Chen, 2003), (Qu et al., 2015), (Koeleman & Koole, 2012), (K. J. Klassen & Yoogalingam, 2014), (K. Klassen & Yoogalingam, 2013), (K. J. Klassen & Yoogalingam, 2009), (Kaandorp & Koole, 2007), (Gupta & Denton, 2008), (Erdogan & Denton, 2013), (Denton & Gupta, 2003), (Cayirli & Gunes, 2014)
Multiple	(W. Y. Wang & Gupta, 2011), (Vanberkel, Boucherie, Hans, Hurink, & Litvak, 2010); (Turkcan et al., 2012); (Drupsteen, van der Vaart, & van Donk, 2013); (Kuiper et al., 2015), (Hulshof et al., 2012), (Gupta & Denton, 2008), (Balasubramanian et al., 2013); (D. L. White, Froehle, & Klassen, 2011); (K. Klassen & Yoogalingam, 2019)

Table 4: Overview of scheduling approaches

Scheduling Approach	References
Online	(K. J. Klassen & Yoogalingam, 2014), (Hulshof et al., 2012), (Gupta & Denton, 2008), (Erdogan & Denton, 2013), (Cayirli & Gunes, 2014), (Balasubramanian et al., 2013); (Samorani & Harris, 2019)
Offline	(Zacharias & Pinedo, 2014), (W. Y. Wang & Gupta, 2011), (L. Robinson & Chen, 2003), (Kuiper et al., 2015), (Erdogan & Denton, 2013), (Denton & Gupta, 2003)

more attention in the literature, because offline systems are easier to model (Ahmadi Javid et al., 2016). Kuiper et al. (2015) compare the online and offline approaches with quadratic and linear loss functions. They observe that the online approach favors the server for both loss functions. Some studies use the results obtained from an offline case to examine the corresponding online case (e.g., Zacharias and Pinedo (2014)).

Choosing the scheduling strategy is not an easy task as both scheduling strategies imply a different model of operations and Appointment System designs for the clinics. When using waiting lists, schedulers should, however, note that patients cannot remain on the list for a long period of time because patient satisfaction will decrease as the urgency level of the patient increases (Buhaug, 2002). Additionally, in an outpatient situation, there is the additional risk that patients will visit the emergency department to be treated sooner (Mayer, Villaire, & Connell, 2005).

2.1.2 Tactical Decisions

Tactical decisions are medium-term decisions related to how patients as a whole are scheduled, or how groups of patients are processed. Outpatient Appointment Systems tactical decisions are covered in the next subsections.

i. Allocation of capacity

It is easily assumed that patients are identical. However, heterogeneity in patient characteristics is an extremely important factor. When patient characteristics are known (such as priority levels, consultation time, disease type, treatment type, and even no-show probabilities) patients should be classified into groups. One of the earlier studies is by Walter (1973), who finds that even a simple grouping of inpatient/outpatient results in substantial improvement in terms of doctors' idle time.

Other studies also commonly recognize the importance of assuming heterogeneous patients and indicate the relevance of allocation (Balasubramanian et al., 2013; Cayirli, Veral, & Rosen, 2006; Cox, Birchall, & Wong, 1985; Deceuninck, Fiems, & De Vuyst, 2018; Qu et al., 2007; Zacharias & Pinedo, 2014). Examples for classifications can be as simple as distinctions of first-time/returning patients and pre-scheduled/same-day patients. Take new/return patients for instance. They often require a longer consultation time compared to returning patients. Deceuninck et al. (2018) set the ratio of the mean consultation time of new patients to the mean consultation time of return patients equal to 1.5. They also present empirical evidence suggesting new patients have higher no-show rates than return patients. Zacharias and Pinedo (2014) also find that the correlation of patient heterogeneity and no-show rates have a significant impact on the optimal schedule and should be taken into consideration.

The same logic should be assumed for service specialties. However, only one work in this scope clustered service types into categories, i.e., Qu et al. (2013). The reason behind this is that these studies focus mainly on appointment scheduling in primary care clinics, in which all services need identical equipment and hence no significant changeover time is incurred between different types of services. Qu et al. (2013) formulate a model to assign service categories to clinic sessions and determine the optimal number of appointments reserved for each service type in each clinic session. The authors' objective was to balance the workload of the providers among clinic sessions, in order to reduce patients' waiting times and providers' idle times.

The consent of all these studies is that the main objective of capacity allocation is dealing with the problem of how the available capacity should be divided among groups. Since classifications can be used for prioritizing, sequencing, scheduling, and adjusting appointment lengths, effectively using them can dramatically reduce costs and waiting times (Deceuninck et al., 2018).

ii. Number of Appointments per consultation session

Studies who consider decisions about the number of appointments per consultation session seek to determine the optimal number of patients that should be scheduled in a consultation session. This number is often calculated in order to minimize the patient waiting time and provider overtime. Vissers (1979), Heaney, Howie, and Porter (1991) and Meza (1998) report a positive relationship between waiting times and the number of appointments in a clinic session (N).

K. J. Klassen and Rohleder (2004) state that since appointment requests come in randomly, some days tend to have higher demand than others and within any given day, some hours are busier than others. The authors also consider that most commonly used options to deal with higher-than-usual demand loads are double-booking clients in a slot and considering server overtime.

Indeed, some clinics accept more patients than their available capacities. Overbooking may allow reduction of the negative impacts of no-shows and to improve patient access (Cayirli & Veral, 2003). However, the number of additional appointments should not be arbitrary. If the level of overbooking is determined inefficiently, it may be ineffective or may lead to longer patient waiting times and system overtime.

Ho and Lau (1992) study finds that the effect of N is mitigated by no-shows and variability of consultation times, as well as other environmental factors that will be discussed below, and thus cannot be easily generalized.

iii. Appointment Rules

Cayirli and Veral (2003) state that the appointment rule used to schedule patients need to be described in terms of a combination of three variables: appointment interval; block-size; begin-block.

An appointment interval is defined as the interval between two successive appointment times (Cayirli & Veral, 2003). These appointment intervals are also referred to as slots and each clinic session is divided into several of these slots, in which patients are scheduled. Appointment intervals can be constant or variable. This is not to be mistaken with consultation (service) time, which is set within the appointment interval, originating either doctor idle time or delays, depending on variability (service times will be addressed below).

Cayirli and Veral (2003) state that a common practice is to set them equal to some function of the mean (and sometimes the standard deviation) of consultation times. The authors also suggest that adjusting appointment intervals proportionally may offset the negative impacts of patient no-shows. Additionally, for unplanned walk-in patients, adjustment requires either leaving open slots or setting appointment intervals relatively longer.

In contrast, several studies show that optimal appointment intervals are not constant, but dome-shaped - meaning they initially increase and then decrease toward the latter part of a session. The considerations are that this improves clinic performance (Denton & Gupta, 2003; Erdogan & Denton, 2013; Kaandorp & Koole, 2007; Kuiper et al., 2015; L. Robinson & Chen, 2003; P. P. Wang, 1993). A modification of this is the plateau-dome (K. J. Klassen & Yoogalingam, 2009) where the central, top portion of the dome is flat. The plateau-dome is best when the realism of an integer restriction is imposed.

In addition to appointment intervals, block-sizes are also taken into account. A block is a set of patients scheduled at the same time, and block size is the number of patients in a block, or the number of patients scheduled at the beginning of a slot within the clinic session (Cayirli & Veral, 2003).

Finally, the begin-block variable - also called the initial block - refers to the number of patients given an identical appointment time at the start of a session.

Table 5 shows several types of common appointment rules.

Table 5: Overview of appointment rule considerations

Appointment Rule	References
Single-block	(Babes & Sarma, 1991); (Barghash & Saleet, 2018)
Individual-block/Fixed-interval	(Cayirli et al., 2006; Cayirli, Veral, & Rosen, 2008); (Wijewickrama & Takakuwa, 2008); (Kaandorp & Koole, 2007)
Individual-block/Fixed-interval with begin-block	(Bailey, 1952); (Ho & Lau, 1992); (Kaandorp & Koole, 2007); (Cayirli et al., 2006, 2008)
Multiple-block/Fixed-interval	(Soriano, 1966); (M. J. B. White & Pike, 1964); (Ho & Lau, 1992); (Cayirli et al., 2006, 2008)
Variable-block/Fixed-interval	(L. Liu & Liu, 1998a)
Individual-block/Variable-interval	(Ho & Lau, 1992); (P. P. Wang, 1993); (L. Robinson & Chen, 2003); (Denton & Gupta, 2003); (Kaandorp & Koole, 2007); (Cayirli et al., 2006)

In a single-block rule (or appointment by date instead of time) patients are assigned to arrive as a block at the beginning of the clinic session (Babes & Sarma, 1991). For example, all patients are scheduled for the same appointment time and they are seen on a first-come, first-served basis. This is the most primitive form of Appointment System, where patients are assigned a date rather than a specific appointment slot. Single-block systems will lead to excessive waiting times for patients, while ensuring that doctors do not stay idle. Barghash and Saleet (2018) consider this rule a common appointment scheduling in

third-world countries.

In individual Appointment Systems, each patient is assigned his own scheduled time. Various other combinations are possible (Wijewickrama & Takakuwa, 2008). Bailey (1952) proposed an individual-block/fixed-interval with an initial block. This is also known as "Bailey's rule" where two patients are scheduled at the start of the session and the rest one by one at fixed intervals. The goal is to keep an inventory of patients so that the doctor's risk of staying idle is minimized if the first patient arrives late or fails to show up.

Researchers have also used multiple block rules (e.g., (Cayirli et al., 2006; Soriano, 1966; M. J. B. White & Pike, 1964), variable block rules with fixed intervals (e.g., (Rising, Baron, & Averill, 1973)), and also variable interval sizes ((Cayirli et al., 2006; Denton & Gupta, 2003; Ho & Lau, 1992; Kaandorp & Koole, 2007; L. Robinson & Chen, 2003; P. P. Wang, 1993)).Take Soriano (1966) for instance. The author studies an appointment system where patients are called two-at-a-time with intervals set equal to twice the mean consultation time (fixed intervals). On the other hand, Kaandorp and Koole (2007) combine a variety of parameters, testing different intervals, in order to find the optimal appointment rule.

Despite all this rules, a consensus among studies is that no single rule performs best in all environments. Yang, Lau, and Quek (1998) propose a universal appointment rule that can be adjusted to perform well in a vast range of clinical environments. Elaborating on the study of Ho and Lau (1992), the rule is expressed as a mathematical function of four environmental parameters (the probability of no-shows, scheduled number of appointments per session, coefficient of variation of service times, and cost ratio of doctor-to-patient time). Once these parameters are estimated, the universal appointment rule can be used to compute the appointment times.

More recently, Cayirli, Yang, and Quek (2012) extend the work of Yang et al. (1998) and suggest a universal Dome rule that combines the advantage of "universality" with dome-shaped appointment intervals. Cayirli and Yang (2014) propose an Appointment System that builds on the universal Dome rule of Cayirli et al. (2012) by including patient classification. The authors compare the performance against the rules previously mentioned and concluded that it performs consistently well under most realistic clinical environments.

2.2 Environmental Factors

Health service providers struggle to balance supply and demand (Gupta & Denton, 2008). Achieving this balance is often difficult on account of variable environmental factors i.e., the uncertainty in the patient arrival and service times, patient and doctors' preferences,

punctuality, cancellations and no-shows, which are described in the section below.

Random Service Times

An early definition of service time (or consultation) can be defined as the sum of all the times a patient is claiming the doctors' attention, preventing him/her from seeing other patients (Bailey, 1952). That being so, it should be considered that the actual time of a consultation may differ from what is expected. Preemptive doctor interruptions (discussed below) can be a reason for this deviations. In practice, doctors may increase their service rate, if only subconsciously, during peak hours knowing that there are many patients waiting (Cayirli & Veral, 2003). High variability of service times deteriorates both the patients' waiting times and the doctor's idle time (Bailey, 1952).

Almost all of the reviewed papers consider uncertainty of service times and assume different types of patients groups. Gupta and Denton (2008) categorize service times as either deterministic or stochastic. Both of these assumptions can include either homogeneous or heterogeneous patient characteristics.

Most analytical studies use Erlang or exponential service times to make their models tractable. (e.g., Kaandorp and Koole (2007); Tang, Yan, and Fung (2014); Turkcan et al. (2012)). Other distributions used to model service time are reviewed in Cayirli and Veral (2003). Kuiper et al. (2015) approximate the service time distributions with phase-type distributions to efficiently determine an optimized schedule with low computational effort.

Qu et al. (2015) state that due to the fact that primary care providers have some control over the service time for each patient, it may be reasonable to assume that service times are deterministic in order to focus on other complicating factors, such as no-shows.

Patients' Preference

Patients usually prefer a specific dates and doctors. These preferences often differ from one patient to another, and may change over time. Nonetheless, as shown in Table 6, most studies do not include patient preferences. Studies like those developed by Babes and Sarma (1991), L. Liu and Liu (1998b) and L. Liu and Liu (1998c) report that some public hospitals do not give appointments for specific patients, sending them to the first available doctor.

Patient preference is first modeled explicitly by Gupta and Wang (2008). In their study, patient choice includes his/her preferred doctor and the convenient time he/she would like to arrive. Patients can switch their choice if the preferred time slot or doctor is not available. W. Y. Wang and Gupta (2011) present a adaptive framework for the design of Outpatient Appointment Systems that dynamically learn and update patient prefer-

Table 6: Overview of environmental factors

Environmental Factors	References
Lateness/Earliness	(K. J. Klassen & Yoogalingam, 2014); (Samorani & Ganguly, 2016); (Tai & Williams, 2012); (K. Klassen & Yoogalingam, 2019)
No-shows	(Kaandorp & Koole, 2007); (K. J. Klassen & Yoogalingam, 2009); (K. J. Klassen & Yoogalingam, 2014); (Koeleman & Koole, 2012); (Zacharias & Pinedo, 2014); (Qu et al., 2007); (Muthuraman & Lawley, 2008); (Samorani & LaGanga, 2015); (W. Y. Wang & Gupta, 2011); (Samorani & Harris, 2019); (Qu et al., 2013)
Random Service Times	(Kaandorp & Koole, 2007); (Denton & Gupta, 2003); (Erdogan & Denton, 2013); (K. J. Klassen & Yoogalingam, 2009); (K. Klassen & Yoogalingam, 2013); (K. J. Klassen & Yoogalingam, 2014); (Koeleman & Koole, 2012); (L. Robinson & Chen, 2003); (Turkcan et al., 2012); (W. Y. Wang & Gupta, 2011)
Lateness/Interruption	(K. J. Klassen & Yoogalingam, 2009); (K. J. Klassen & Yoogalingam, 2014); (K. Klassen & Yoogalingam, 2013); (Vissers, 1979); (Rising et al., 1973); (M. J. B. White & Pike, 1964); (L. Liu & Liu, 1998b); (L. Liu & Liu, 1998c)
Patients' Preference	(Gupta & Denton, 2008); (W. Y. Wang & Gupta, 2011)

ences. The authors state that optimal access policies - that are well prepared for matching randomly arriving patients' appointment requests - only happen if patient preferences are tracked. It is only logical that patient satisfaction increases with a well prepared Appointment System.

More recently, N. Liu, Finkelstein, Kruk, and Rosenthal (2017) provide a study that demonstrates how an Appointment System performance may affect patient satisfaction, offering insights on how to use forecasting methods to improve patient experience of care and examining the effect of individual-difference variables, such as gender. The authors observe an interesting effect with respect to how patients trade off waiting times and satisfaction. Gupta and Denton (2008) also state that doctors also have preferences but no paper in this literature considered it.

Lateness/Earliness of patients

Patient punctuality is defined as the difference in time between patients' arrival for an appointment and the scheduled time of said appointment (M. J. B. White & Pike, 1964). Thus, punctuality includes patient earliness and/or lateness. Empirical evidence suggests patients generally do not arrive on time, (Cayirli et al., 2006; Lehaney, Clarke, & Paul, 1999; Vissers, 1979; M. J. B. White & Pike, 1964).

Although patient arriving early is just as impactful as arriving late for Outpatient Appointment Systems planning and appointment scheduling, studies that include this factor tend to focus on lateness (K. J. Klassen & Yoogalingam, 2014). Nonetheless, only a few papers consider it. Koeleman and Koole (2012) noted that most papers on the design of appointment scheduling systems assume that a patient is punctual, meaning lateness is not considered. Ahmadi Javid et al. (2016) state that this happens due to the complexity caused by considering patient lateness in a mathematical model.

Tai and Williams (2012) model patient earliness and lateness as a combination of some common distributions (normal and lognormal distributions) and their modified forms with consideration of various patient behaviour patterns. Some studies also use empirical data to approximate this distribution, for example, normal distribution (Cayirli et al., 2006, 2008), or exponential distribution (Cox et al., 1985).

One issue that clinics face is how to deal with patients arriving out of the scheduled order. Samorani and Ganguly (2016) studied the problem of whether an available provider should see an early patient right away (preempt) or wait for the next scheduled patient. Based on simulation results, the authors reported the conditions under which the policy outperforms the always-preempt policy (e.g., high-service-level clinics with low variability in patient lateness, long service duration's, and when patients tend to arrive early rather than late).

Several simulation studies report that patient earliness or lateness are highly significant factors influencing the performance of appointment scheduling systems (e.g., (K. J. Klassen & Yoogalingam, 2014), (Cayirli et al., 2006), (Cayirli et al., 2008) and (Zhu, Chen, Leung, & Liu, 2017)).

Lateness and Interruption Level of Doctors

Doctors' lateness can lead to a reduction in clinic efficiency by extending patient waiting times (K. Klassen & Yoogalingam, 2013). This factor is measured as lateness to first appointment.

There is agreement among all studies that patient waiting times are highly sensitive to doctors' lateness. If the doctor does not start the appointment on time, a delay factor builds up from the start to finish of the clinic session.

On the other hand, there are doctors' interruptions (also called "gap-times" in (Cayirli

& Veral, 2003)). There are two main types of interruptions: non-preemptive and preemptive (Ahmadi Javid et al., 2016). Non-preemptive interruptions occur between patient consultations (e.g., (K. Klassen & Yoogalingam, 2013)). Preemptive interruptions occur during patient consultations (e.g., (Krishnamoorthy, Pramod, & Chakravarthy, 2014)) and can be a factor in service times variability.

Although this environmental factor has a significant impact on the reliability and performance of an Outpatient Appointment System, it has received limited attention in the literature.

No-shows

One of the major problems that most outpatient clinics are confronted with is patient no-shows.

According to W. Y. Wang and Gupta (2011), based on factors affecting the no-show probability, no-shows can be divided into five categories: homogeneous, wait-dependent, patient dependent, service-dependent, and time-dependent.

Furthermore, a number of articles outside the Operational Research literature investigate patient no-shows empirically. For example, see Oppenheim, Bergman, and English (1979), Pesata, Pallija, and Webb (1999), Moore, Wilson-Witherspoon, and Probst (2001) and Gallucci, Swartz, and Hackerman (2005). All of these point to patient no-shows as being a significant problem in appointment scheduling and find that no-show rates also depend on a variety of factors including race, gender, socioeconomic status, physical restraints and other external factors. In particular, Gallucci et al. (2005) conclude that no-show and cancellation rates increase with appointment delays and observed waiting times.

Dantas, Fleck, Cyrino Oliveira, and Hamacher (2018) summarize the findings in the literature dealing with no-shows in appointment scheduling. The authors conclude that average no-show rate across all studies observed by them was found to be 23.0%, and further analysis revealed that this rate was highest in the African continent (43.0%) and lowest in Oceania (13.2%). Additionally, in the majority of the studies survey by them, the most important factors to influence no-show were found to be lead time (time interval between the date when the appointment is registered in the clinic's scheduling system and the actual appointment date) and prior no-show history. They also identify patient characteristics that were more frequently associated with no-show behavior as well as external factors leading towards them.

Two strategies are proposed in the literature to manage the negative effects of no-shows: reducing appointment intervals and overbooking (Ahmadi Javid et al., 2016), (K. Klassen & Yoogalingam, 2013), (Samorani & LaGanga, 2015). In fact, Samorani and LaGanga (2015) state that when overbooking, while it is desirable to maximize the

number of patients seen, it is also desirable to limit patient waiting time and clinic overtime. To solve this trade-off, the authors developed a dynamic scheduling procedure which considers no-show predictions. These no-show predictions might be a way to improve a clinics' performance.

2.3 Solution Methods and Modeling Approaches

In this section, modeling approaches and solution methods within the Outpatient Appointment Systems literature are discussed. Modeling approaches concern techniques to translate real case problems into mathematical formulations, while solution approaches refer to systematic methods to solve the referred problems.

2.3.1 Modeling Approaches

In terms of modeling, most approaches make use of either stochastic optimization and dynamic programming or deterministic models.

Stochastic programming allows the user to minimize or maximize an objective function in the presence of randomness. For instance, simulation optimization is a stochastic optimization method that enables a search for solutions in problems where some or all of the system parameters are stochastic (e.g., (K. Klassen & Yoogalingam, 2013, 2019; K. J. Klassen & Yoogalingam, 2009, 2014)). It is well suited for problems where uncertain parameters can be represented by probability distribution functions. The problem formulation for simulation optimization algorithms specifies the objective function and constraints as a set of discrete-event simulation models in which a heuristic guides the search for an optimum.

Considering deterministic models on the other hand, often involve (mixed) Integer Linear Programming. Deterministic models are often used to formulate problems at the tactical level that are less affected by the uncertainty caused by random arrivals and random service times. These models are also used in specialty clinics (such as radiotherapy or cardiology clinics) where service time in each stage of the treatment procedure is deterministic and no-show probability is close to zero. Typical complications for these models include capacity and due date constraints with multi-resource, multi-stage treatment procedures (Conforti, Guerriero, & Guido, 2008, 2010, 2011; Pérez, Ntaimo, Wilhelm, Bailey, & McCormack, 2011; Qu et al., 2013; Turkcan et al., 2012).

2.3.2 Solution Methods

In general, two broad classes of solution methods regarding research on appointment scheduling exist as follows: analytical studies and simulation studies.

Analytical studies employ theoretical tools, such as queuing theory, mathematical programming and dynamic programming, to determine the number of patients for each appointment interval (slot, session or day) or the length of such interval. Heuristic approaches are also commonly applied in addressing problems for which no analytical solution method exists, but where analytical formulations of the objective function and constraints do exist. Although these methods do not guarantee optimal solutions, sometimes they are indispensable in practical because of the presence of complex environmental characteristics and stochastic factors (e.g., (Balasubramanian et al., 2013; L. Robinson & Chen, 2003)).

Meanwhile, simulation studies focus on specific appointment scheduling systems in complex environments, with the goal of comparing among various systems and measuring the effect of key factors. This can be seen as an advantage (Cayirli & Veral, 2003). In fact, studies conduct simulation experiments to evaluate the performance of alternative Appointment Systems and/or understand the relationship between various environmental factors and various performance measures.

2.4 Considerations

There are many approaches that can be followed to model an appointment system scheduling problem. Those approaches differ mainly on the type of decisions to be made, the patients considered, the conditions of the hospital under study and the objectives to be achieved.

Strategic decisions have a significant effect on modeling and on the practical applicability of the presented solution methods. Such decisions are most frequently treated as inputs into an Outpatient Appointment Systems model, and only a few optimization studies are found that compare strategic-level options based on numerical experiments.

The problems addressed at the tactical level, on the other hand, aim to determine the Outpatient Appointment System structure. While it seems that determining the optimal level for tactical decisions can significantly increase system performance in the long term, some of these decisions have only received limited attention in the Operational Research literature (e,g., appointment scheduling window; fairness; workload balance; demand forecast; congestion; patient and consultation characterization and allocation). Including these decisions within the objective functions can lead to lower patient waiting time and minimize provider idle and overtime by indirectly affecting the clinics organization.

In addition to that, it is found that optimization studies sometimes neglect environmental factors that may further complicate the mathematical models (e.g., patient and physician preferences; no-shows). As has been noted repeatedly in the previous studies,

these factors appear in realistic problems, and hence considering them can help increase the applicability of the resulting models.

CHAPTER III - MODELS DEVELOPMENT

In this chapter, the models proposed to formulate the appointment scheduling problem are presented as Integer Linear Programming models. The models are inspired by Qu et al. (2013), who approach their decision-making problem with the objective of balancing provider workload among clinic sessions, by categorizing service types and assigning exactly one service category to each session, to determine the optimal number of appointments that should be reserved for each service type in the clinic session.

The suggested methodology approached towards the Appointment System problem is defined in Section 3.1. The Integer Linear Programming models are described in therms of decision and objective in Section 3.2, the notation used to describe the objective function and constraints is presented in section 3.3. Section 3.4 and Section 3.5 present Model I and Model II formulation, respectively. This chapter winds up with conclusions and considerations towards the modeling process in section 3.6.

3.1 Methodology

Overall, this problem aims to control waiting times, minimize idle times of resources, make hospitals more cost-efficient, provide fairness to patients and doctors and ultimately, provide common satisfaction to all sources. As previously mentioned, this can be done by applying an appropriate, sensitive, and responsive procedures.

This dissertation contributions will be on the tactical level. The potential effectiveness of the approached methodology is demonstrated by its application towards different scenarios of testing. In order to do this, different numerical examples will be generated.

The testing will approach different levels of appointment demands within a generated population, towards clinical specialties. Assuming the maximum capacity of resources that a certain hospital can observe, considering different levels of demands allows the understanding of the workload limits from clinical sessions, rooms and workload limits for doctors'.

As previously mentioned, including tactical decisions within the objective functions can lead to lower patient waiting time and minimize provider idle and overtime by indirectly affecting the clinics organization.

Clinics can identify with this methodology either for available specialties, distinguished by their expected service times, and the number of rooms which varies depending on the capacity of the clinics.As an example, specialties considered could be Cardiology, Neurology, Urology, Oncology and many others. The assumption on which specialty should be tested should be made by the clinics or hospitals based on their average appointment times for each specialty. Additionally, available resources can also be combined with

the number of rooms available.

Appointment times were considered deterministic in order to focus on other decision variables, and dependent of patient categorizations and the different types of medical specialties considered (i.e., first-time patients have higher values of service times than return patients; and, different specialties have different expected service times).

3.2 Description

Similarly to Qu et al. (2013), the objective is to balance workload fairness.

Although, whilst Qu et al. (2013) define a session as a time period during which providers see a series of scheduled patients without a break, within this case a session represents a day of scheduling and it can contain multiple appointment offices, referred to as rooms, having exactly one service category (i.e., specialty) assigned to it.

Therefore, as Qu et al. (2013) determine the optimal number of appointments that should be reserved for each service type in a clinic session, this models determine the optimal number of appointments that should be reserved for each patient type in a room assigned to a specialty.

This session definition is important in the understanding that Qu et al. (2013) balance the workload difference among sessions whilst the presented approaches minimize the workload differences among rooms, within a session.

The reason choose to balance the workload is that even though an optimal decision can be made independently for each room, unbalanced workload may lead to long patient waiting time and doctor overtime for rooms with higher workloads and lead to long doctor idle time for rooms with lower workloads.

With this objective in mind, the development is divided in two different modeling approaches, referred to as Model I and Model II.

Model I, formulated in Section 3.4, has the main objective of minimizing the total workload differences among rooms. On the other hand, Model II, formulated in Section 3.5, proposes a new objective function that minimizes the maximum workload difference, with a $minxmax$ decision process. Both Model I and Model II also categorize patient types as patient classifications that influence service times.

3.3 Notation

This section approaches the notation used towars the formulation of the Integer Linear Programming models. The indices, sets and subsets are summarized on Table 7 while parameters are presented in Table 8. Decision variables and auxiliary variables are described in Table 9.

Table 7: Indices, sets and subsets

Indices, sets and subsets	
s and $s' \in S$	Room
$c \in C$	Specialty
$t \in T$	Service type

Table 8: Parameters

Parameters	
η_s	Duration for which room s is available
δ_{ct}	Duration of appointment of specialty c and type t
λ_{ct}	Demand of appointment of specialty c and type t
M	Big number

Let S be the set of rooms, C be the set of specialties and T be the set of service types.

In addition, parameters η_s, δ_{ct} and λ_{ct} denote the expected duration of room s and the expected duration and demand defined for appointment of specialty c and type t, respectively.

Table 9: Decision and Auxiliary Variables

Decision Variables	
$w_{s,s'}$	Workload difference between any two rooms s and s'
Auxiliary Variables	
x_{cs}	1, if specialty c is assigned to room s; 0, otherwise
y_{cts}	Number of appointments of specialty c and type t that should be planned for room s
w_s	Workload from room s in minutes

The first set of decision variables, $w_{s,s'}$, is defined as the difference of the workload between any two rooms s and s'. In account of this, auxiliary variables w_s are computed as the workload on room s, in minutes. Moreover, auxiliary variables y_{cts} account the number of appointments of specialty c and type t that should be planned for room s. Variables x_{cs} assume value 1 if specialty c is assigned to room s, and 0 otherwise.

3.4 Model I Formulation

In this section, the objective function and constraints of the Integer Linear Programming Model I are explained, using the indices, sets, subsets, parameters and decision variables previously described in this chapter. The objective function is presented in Expression (1). Constraints are formulated with Expressions (2) to (10).

$$\min \quad \sum_{s,s' \in S,\, s \neq s'} \frac{1}{2}\, w_{s,s'} \tag{1}$$

$$\text{s.t.:} \quad \sum_{c \in C} x_{cs} = 1, \qquad \forall\, s \in S \tag{2}$$

$$\sum_{s \in S} y_{cts} = \lambda_{ct}, \qquad \forall\, c \in C, \forall\, t \in T \tag{3}$$

$$\sum_{c \in C} \sum_{t \in T} y_{cts} \times \delta_{ct} = w_s, \qquad \forall\, s \in S \tag{4}$$

$$w_{s,s'} = \left| w_s - w_{s'} \right|, \qquad \forall\, s, s' \in S,\, s \neq s' \tag{5}$$

$$w_s \leq \eta_s \qquad \forall\, s \in S \tag{6}$$

$$\sum_{t \in T} y_{cts} \leq M \times x_{cs}, \qquad \forall\, c \in C, \forall\, s \in S \tag{7}$$

$$\sum_{t \in T} y_{cts} \geq x_{cs}, \qquad \forall\, c \in C, \forall\, s \in S \tag{8}$$

$$w_{s,s'}, y_{cts}, w_s \in \mathbb{R}_0^+ \tag{9}$$

$$x_{sc} \in [0, 1] \tag{10}$$

Expression 1 intends to minimize the total workload differences among rooms, and therefore, balance the workload. Constraints 2 guarantee that a room s is only assigned with one specialty c. For each type of service, Constraints 3 enforce that the appointments reserved must meet all the demand for the service type t and specialty c. The expected total service time of a room s appointments equals the room workload, as defined in Expression 4. Constraints 5 define the difference among room workloads [1]. Since every room must be limited, Constraints 6 were expressed. Additionally, two sets of linking constraints are required to connect variables y_{cts} and x_{cs} - Expressions 7 and 8. Expressions 9 and 10 state the variables domain.

[1] Linearized format: $w_s - w_{s'} \leq w_{s,s'} \quad and \quad w_{s'} - w_s \leq w_{s,s'}$

3.5 Model II Formulation

In this section, the adaptations made to generate Model II are explained. The new objective function is presented in Expression (11). A new constraint is formulated with Expression (12).

$$\min \quad w \tag{11}$$

$$\text{s.t.:} \quad w_{s,s'} \leq w, \qquad \forall\, s, s' \in S \tag{12}$$

In Model II, a new variable is created, denoted by w which represents the maximum value of the workload difference between any two rooms s and s', or $w_{s,s'}$. The new objective function minimizes the maximum workload difference, with a $minxmax$ decision process. This is defined in Expression 11. Expression 12 defines the new variable.

3.6 Considerations

The objective on Model I, as seen, is to minimize the workload differences among rooms. As a result output, Model I provides all differences between any two or more rooms being minimized (i.e., results for workload balance).

On the other hand, a $minxmax$ problem seeks to minimize the maximum value of a number of decision variables. It is sometimes applied to minimize the possible loss for a worst case (maximum loss) scenario.

In this case, Model II represents a $minxmax$ decision process that intents to minimize the maximum value of the workload differences between any two or more rooms instead of the aforementioned Model I, that intents to do it so for every difference generated.

The output for Model II suggests that the higher value assumed for the difference is the one being minimized. That being, for comparison measures, Model II objective function results are to be compared with the average and higher value obtained for workload differences on Model I outputs, since the objective function, in this case, represents the total sum of the differences generated. The highest workload difference observed is to be referred to as the W_{max}.

CHAPTER IV - NUMERICAL EXAMPLES

The formulation described in Chapter IV demands a specific type of input. This chapter is organized in two sections: Section 4.1 describes the inputs for generation; Section 4.2 explains the main characteristics of the input generation.

4.1 Description

To validate and test the quality of the solutions, it is necessary to have input data. The models are structured and parameterized according to randomly generated data.

For this, a programming model was coded in Java to generate said random numerical examples, that reasonably represent the rooms available in each session, the number of specialties, duration for which a room is available and service times, as well as demand for each type of patient within each specialty. A summary of the numerical examples' generator inputs is presented in Figure 1. The respective outputs are the results obtained for the workload balance.

Figure 1: Inputs for numerical examples' generation

Input for no. of rooms, no. of specialties and no. of service types.	Input for room availability (maximum duration)	Input for appointment duration (for each service type within each specialty)	Input for demand duration (for each service type within each specialty)	Results for workload balance

4.2 Generation

Numerical examples are required for the computational experiments, including model validation and sensitivity analysis. Although it was not possible to obtain real data, a range of default guideline values was assumed based on time dimensions and the literature reviewed. This range of randomized values makes it possible for hospitals to assume the values that identify with their realities, and apply them to their practices. This information is presented in this section.

There are some characteristics that need to be capped values based on feasibility and realistic assumptions of hospital data.

For problem complexity simplification, there is a limited number of appointment rooms available in each session to be used by the hospitals that is necessary to establish. Additionally, the number of specialties generated has to be established based on the number of rooms available.

In terms of demand, the values had to be generated based on specialty and type of patient. The type of patient, in this case, represents any assumption that can distinguish two appointment times based on a previous characterization prior to the schedule.

Moreover, based on the problem definition, it is necessary to define the number of minutes in which a room can work.

These parameters are summarized in Table 10.

Table 10: Parameters for numerical examples' generation procedure

Parameters for numerical examples' generation procedure	
Big Number	$M = 10000$
Rooms	$S \in [1, 15]$
Specialty	$C \in [1, S]$
Service type	$T = 2$
Duration for which room s is available	$\eta_s = 30(1 + x)$, $x \in [0, 15]$
Duration of appointment of specialty c and type t	$\delta_{ct} \in [10, 30]$
Demand of appointment of specialty c and type t	$\lambda_{ct} \in [0, 15]$

Instead of generating unlimited rooms, every numerical example was limited to a maximum of 15 rooms. Naturally, the number of specialties generated was between 1 and the number of rooms, since it is not possible to have more specialties than rooms available.

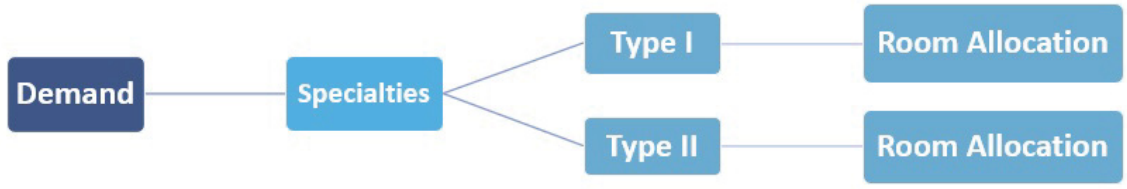
For service type, every numerical example assumed 2 types of patient distinguished by appointment times. The type of patient is subjective, but in this case it can be considered as any patient characterization that defines one greater than the other.

Considering that appointments are usually performed between a defined time by hospitals or clinics, a set time had to be established. That being, a maximum of 8 hours (i.e. 480 minutes) of appointments can be scheduled in each room, representing a workday. Accordingly, the appointment times were capped in a gap of 10 to 30 minutes and every room could assume a maximum of 16 slots, with 30 minutes each, totaling the mentioned 8 hours of maximum use of a room.

Lastly, the demand generated assumed values between 0 to 15 patients for each type within each specialty.

Figure 2 shows the allocation process for these models, assumed by every numerical example generated.

Figure 2: Allocation process for numerical examples' generation



CHAPTER V - RESULTS AND ANALYSIS

In this chapter, the models are validated through 30 numerical examples each (Section 5.1).

Moreover, these numerical examples are used to analyze the dimensions of the problem and sensitivity analyses on changes in parameters and to the objective function (Section 5.2). The chapter ends in Section 5.3 with some considerations about the results for each Model.

5.1 Model Validation

The experimental methodology for this study was implemented on a AMD Ryzen™ 5 3600 6-Core CPU (3.6GHz-4.2GHz) and 16 GB RAM computer, and the Windows 10 operating system. The models were implemented in Java with Eclipse Java Mars.2, using the callable library of ILOG CPLEX 12.8.0. CPLEX is a powerful tool to model and solve optimization problems. It uses a branch-and-cut method to solve the integer programming models ("IBM Knowledge Center," n.d.). Tests are performed in the 60 numerical examples described in the previous chapter.

5.2 Sensitivity Analysis

In this section, the sensitivity of both Models to the changes in parameters, numerical examples and results are analyzed and compared. The impact of changes in the quality of the solutions obtained is also evaluated. All numerical examples were used to test the sensitivity of the models to the dimensions of the problem and its complexity. Changes to rooms available, number of specialties, demand and average appointment time are tested and compared in both Models, regarding model run time and workload balance.

Firstly, model run time is observed. Figure 3 and Figure 4 display the model run time by rooms available for Model I and II, respectively. Regarding the number of specialties in each numerical example, run times can be observed in Figure 5 and Figure 6.

Figure 3: Model I Run Time (seconds) by rooms available

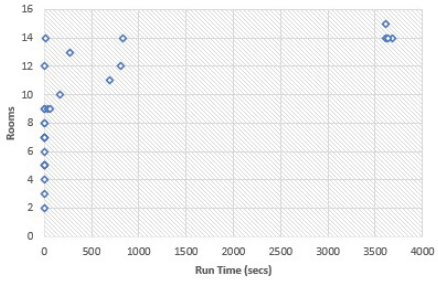


Figure 4: Model II Run Time (seconds) by rooms available

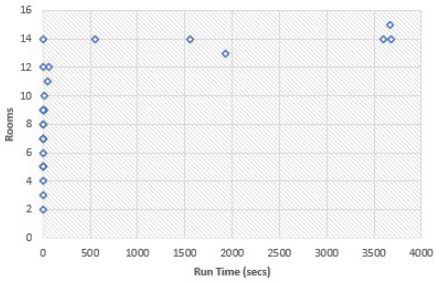


Figure 5: Model I Run Time (seconds) by number of specialties

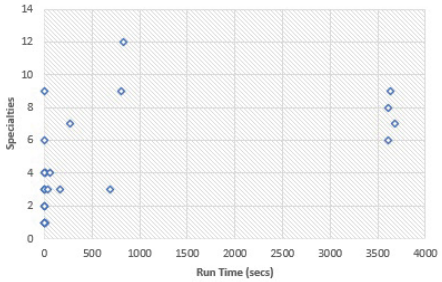
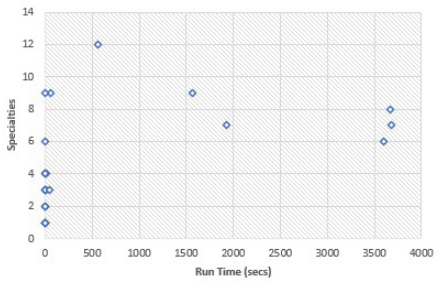


Figure 6: Model II Run Time (seconds) by number of specialties



The models show good results in reasonable run times for numerical examples with less than approximately 10 rooms available. Higher run times are observed when numerical examples surpass these number of available rooms. Although these results are fairly close for both Models, Model II shows more frequently reasonable run times when the number of rooms is high. A similar behaviour can be observed for the model run time by the number of specialties.

Concerning the total demand for each type and specialty, results show a similar behaviour for model run time in both Models. Short run times are more common when the total daily demand generated does not exceed approximately 100. When the generated demand values increase, it becomes harder to define a correlation between the variables. This can be observed in Figure 7 and Figure 8.

Figure 7: Model I Run Time (seconds) by demand

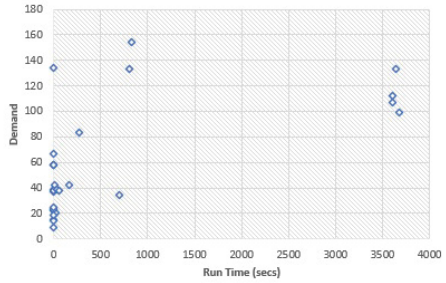
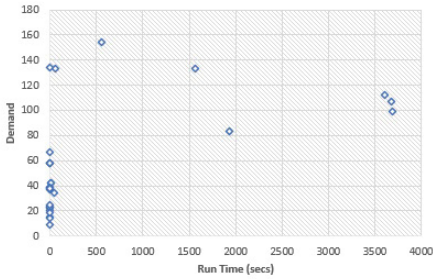


Figure 8: Model II Run Time (seconds) by demand



Finally, regarding appointment duration, the results shown suggest that there is a lack of relation between the model run time and the average appointment duration's generated.

Figure 9: Model I Run Time (seconds) by average appointment duration (minutes)

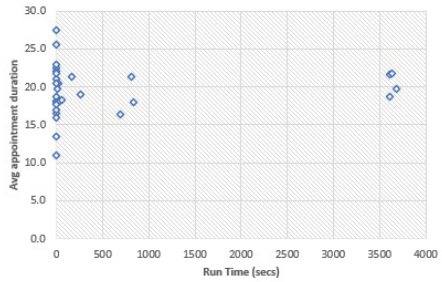
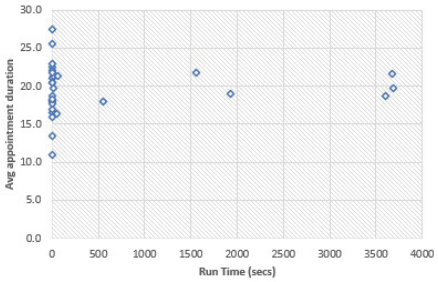


Figure 10: Model II Run Time (seconds) by average appointment duration (minutes)



With reference to workload balance, Figure 11 and Figure 12 show that the number of rooms do not affect directly the workload differences although, on the other hand, changes in the number of specialties are directly related with the increase in the differences observed, as shown in Figure 13 and Figure 14.

Figure 11: Model I average workload differences by rooms available

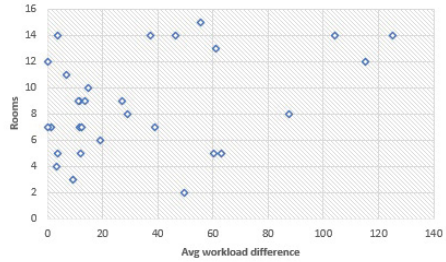


Figure 12: Model II average workload differences by rooms available

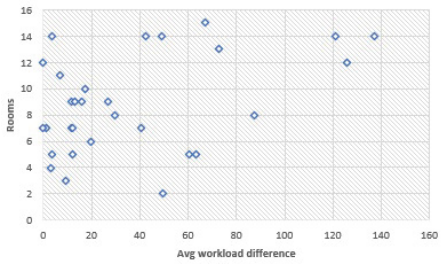


Figure 13: Model I average workload differences by number of specialties

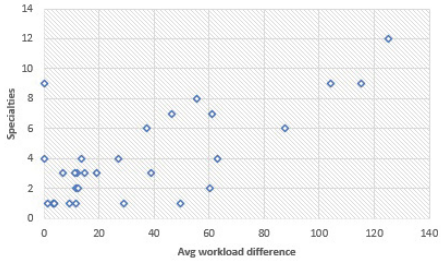


Figure 14: Model II average workload differences by number of specialties

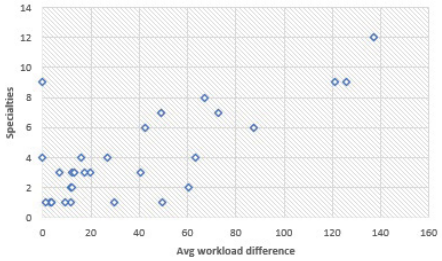


Figure 15 and Figure 16 show that an increase in demand is also related to an increase in the workload differences, such as a decrease in demand is related to a decrease in workload differences.

Although the number of rooms alone does not affect the average workload differences observed, changes in the number of specialties available do affect the workload differences and thus, the total demand observed and allocated to each specialty, will also have an impact.

Figure 15: Model I average workload differences by demand

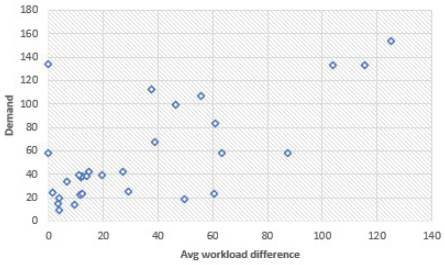
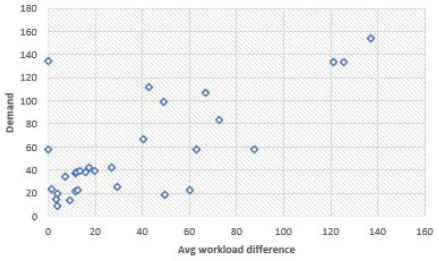


Figure 16: Model II average workload differences by demand



Furthermore, the results shown in Figure 17 and Figure 18 provide the information that, once again, the average appointment duration does not seem to be directly related with an increase in the workload balance.

Figure 17: Model I average workload differences by average appointment duration (minutes)

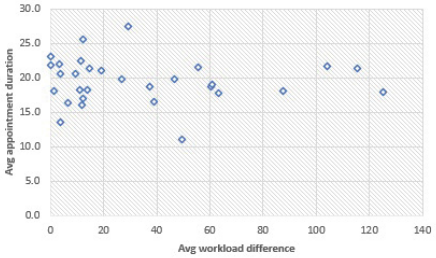
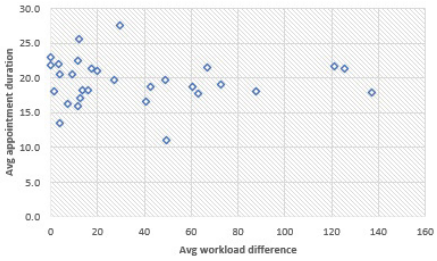


Figure 18: Model II average workload differences by average appointment duration (minutes)



For better understanding of the results on workload balance, the following Figures 19, 20, 21 and 22 show the relation of the maximum value for w observed on each Model (W_{max}), with their positively related variables (i.e. number of specialties and demand).

Figure 19: Model I W_{max} by number of specialties

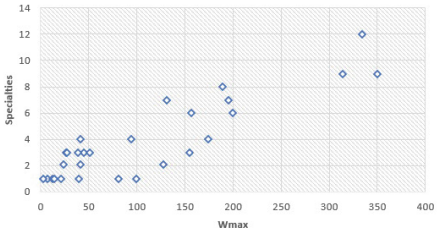


Figure 20: Model II W_{max} by number of specialties

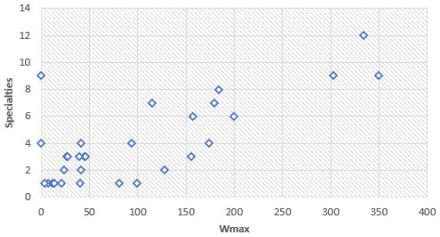


Figure 21: Model I W_{max} by demand

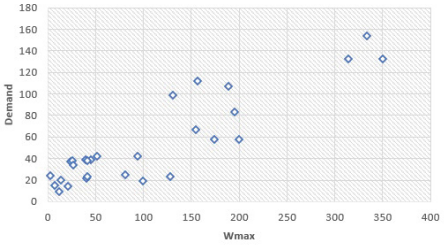
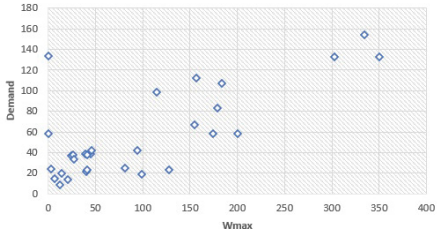


Figure 22: Model II W_{max} by demand



This results show that the models take more impact to run time and the objective function itself (i.e. the workload balance) with changes to either the number of clinical specialties available on each numerical example and the total demand for said specialties and service types.

5.3 Considerations

The number of constraints and the number of variables generated with the models for each one of the 60 numerical examples in 1-hour tests are described accordingly to number of rooms available, number of specialties, total demand for each type and specialty, average appointment duration in minutes, average workload difference and W_{max}, in Table 11 and 12, for Models I and II respectively.

Table 11: Model I results overview

Numerical example	No. of Rooms	No. of Specialties	Total Demand	Avg. Appointment Duration (minutes)	Avg. Workload difference (minutes)	W_{max} (minutes)	No. of Variables	No. of Constraints
1	6	3	39	21	19.3	45	96	120
2	9	1	22	22.5	11.56	40	117	191
3	8	6	58	18.08	87.59	200	216	244
4	4	1	15	22	3.38	7	32	46
5	15	8	107	21.56	55.68	189	600	721
6	5	2	23	18.75	60.32	128	60	79
7	5	1	9	13.5	3.84	12	45	67
8	7	3	67	16.5	38.94	155	119	153
9	14	1	20	20.5	3.67	14	252	436
10	7	1	24	18	1.47	3	77	121
11	14	7	99	19.71	46.52	131	504	616
12	13	7	83	19	60.92	195	455	547
13	9	4	42	19.75	26.96	94	198	251
14	3	1	14	20.5	9.33	21	21	29
15	10	3	42	21.33	14.58	51	200	276
16	7	2	37	16	11.67	24	98	137
17	7	2	23	17	12.33	41	98	137
18	14	12	154	17.96	125.33	334	714	766
19	5	3	38	25.5	12.16	26	75	91
20	12	9	133	21.33	115.35	350	480	534
21	9	3	39	18.16	11.21	39	171	231
22	12	9	134	21.83	0	n/a	n/a	n/a
23	7	4	58	23	0	n/a	n/a	n/a
24	5	4	58	17.75	63.2	174	90	103
25	2	1	19	11	49.5	99	12	16
26	8	1	25	27.5	29.16	81	96	154
27	14	6	112	18.75	37.52	157	462	586
28	14	9	133	21.72	104.03	314	588	676
29	11	3	34	16.33	6.71	27	231	325
30	9	4	38	18.25	13.72	41	198	251

For Model I, regarding the number of variables, the number of rooms available is the set with greater impact. As for the number of constraints, the number of rooms and the number of specialties are the inputs with greater influence to the results observed.

The number of variables for Model II increases by 1 since w was introduced. As to the number of constraints, we can observe an increase of $|s| \times |s|$ considering the addition of Expression 12 to the initial model, for the same problem defined.

Table 12: Model II results overview

Numerical example	No. of Rooms	No. of Specialties	Total Demand	Avg. Appointment Duration (minutes)	Avg. Workload difference (minutes)	W_{max} (minutes)	No. of Variables	No. of Constraints
1	6	3	39	21	19.78	45	97	156
2	9	1	22	22.5	11.6	40	118	272
3	8	6	58	18.08	87.59	200	217	308
4	4	1	15	22	3.38	7	33	62
5	15	8	107	21.56	67	184	601	946
6	5	2	23	18.75	60.32	128	61	104
7	5	1	9	13.5	3.84	12	46	92
8	7	3	67	16.5	40.57	155	120	202
9	14	1	20	20.5	3.67	14	253	623
10	7	1	24	18	1.47	3	78	170
11	14	7	99	19.71	48.99	115	505	812
12	13	7	83	19	72.64	179	456	716
13	9	4	42	19.75	26.96	94	199	332
14	3	1	14	20.5	9.33	21	22	38
15	10	3	42	21.33	17.18	46	201	376
16	7	2	37	16	11.67	24	99	186
17	7	2	23	17	12.33	41	99	186
18	14	12	154	17.963	137.18	334	715	962
19	5	3	38	25.5	12.16	26	76	116
20	12	9	133	21.33	125.58	350	481	678
21	9	3	39	18.17	13.33	39	172	312
22	12	9	134	21.83	0	n/a	n/a	n/a
23	7	4	58	23	0	n/a	n/a	n/a
24	5	4	58	17.75	63.2	174	91	128
25	2	1	19	11	49.5	99	13	20
26	8	1	25	27.5	29.44	81	97	218
27	14	6	112	18.75	42.61	157	463	782
28	14	9	133	21.72	120.93	303	589	872
29	11	3	34	16.33	7.11	27	232	446
30	9	4	38	18.25	15.95	41	199	332

The number of constraints and the number of variables generated with the models and the relative gap and model run time for each one of the numerical examples generated are described in Table 13 for both Models.

The relative gap is calculated as the percentage that the difference between the best bound and the best integer solution represents when compared to the best integer solution value [2].

Table 13: Results for the number of variables, number of constraints and relative gap (%)

Numerical	No. of Variables		No. of Constraints		Relative Gap (%)		Run Time (seconds)	
Example	Model I	Model II	Model I	Model II	Model I	Model II	Model I	Model II
1	96	97	120	156	0	0	1	0
2	117	118	191	272	0	0	1	0
3	216	217	244	308	0	0	2	1
4	32	33	46	62	0	0	0	0
5	600	601	721	946	18.41	20.47	3606	3672
6	60	61	79	104	0	0	0	0
7	45	46	67	92	0	0	0	0
8	119	120	153	202	0	0	0	0
9	252	253	436	623	0	0	18	0
10	77	78	121	170	0	0	0	0
11	504	505	616	812	12.53	18.26	3679	3683
12	455	456	547	716	0	0	267	1929
13	198	199	251	332	0	0	8	9
14	21	22	29	38	0	0	0	0
15	200	201	276	376	0	0	165	9
16	98	99	137	186	0	0	1	1
17	98	99	137	186	0	0	0	0
18	714	715	766	962	0	0	835	557
19	75	76	91	116	0	0	0	0
20	480	481	534	678	0	0	809	60
21	171	172	231	312	0	0	40	1
22	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
23	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
24	90	91	103	128	0	0	0	0
25	12	13	16	20	0	0	0	0
26	96	97	154	218	0	0	1	0
27	462	463	586	782	13.63	5.10	3609	3601
28	588	589	676	872	5.50	0	3636	1561
29	231	232	325	446	0	0	696	47
30	198	199	251	332	0	0	58	1

[2] $Relative\ Gap = \frac{(Best\ Bound - Best\ Integer)}{Best\ Integer}$

Considering 30 numerical examples for each Model, with the referred dimensions, the problems generated are to be solved with 12 to 714 variables for Model I, 13 to 715 variables for Model II, 16 to 766 constraints for Model I and 20 to 962 constraints for Model II. Numerical examples 22 and 23 were not solved in 1-hour tests.

The worst result for both Models is obtained for numerical example 5, with a relative gap of 18.41% and 20.47%, respectively. Moreover, numerical examples 11, 27 and 28 did not reach a 0% gap in 1-hour tests. Table 11 and Table 12 show that these numerical examples have in common a high number of specialties and total demand.

In terms comparison between Models, it is possible to observe that Model II had one more optimal solution and an overall shorter aggregate model run time. On the other hand, regarding average workload differences and W_{max} both Models show fairly similar results, as shown in Figure 23 and Figure 24.

Figure 23: Average workload difference for Model I and Model II

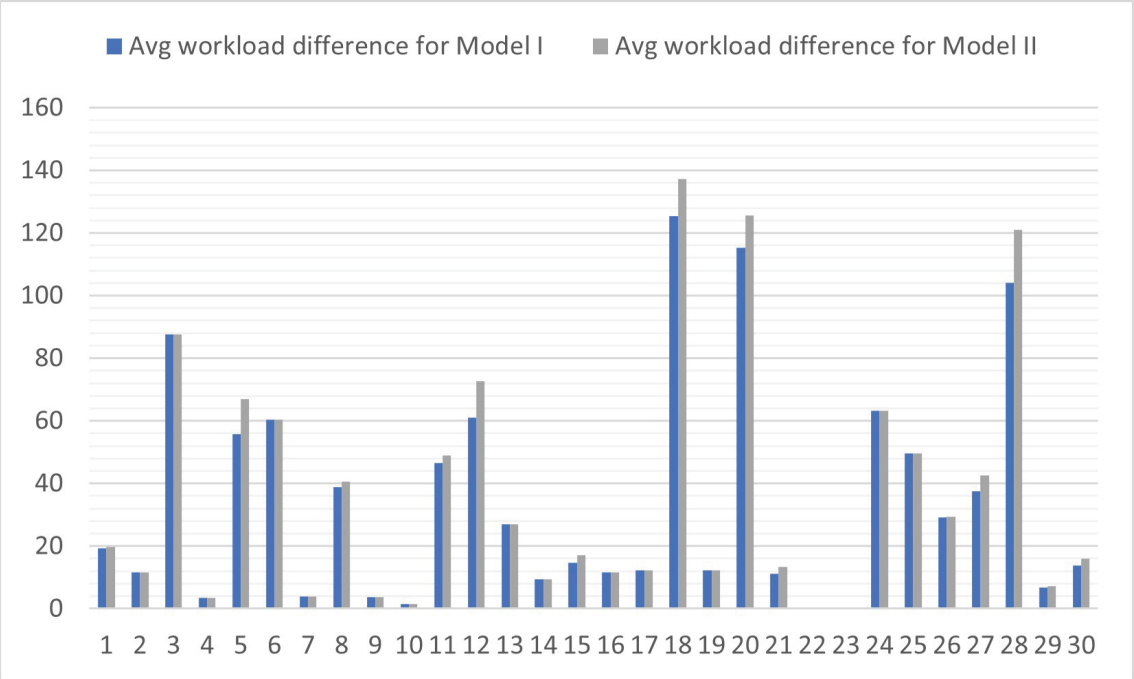
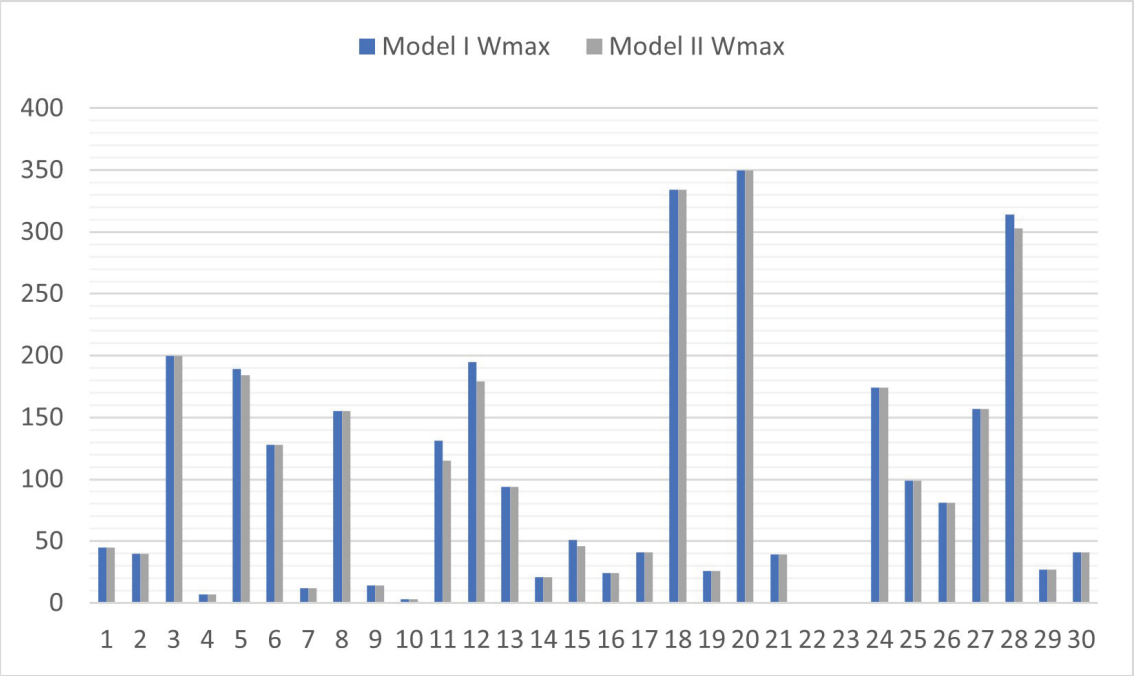


Figure 24: W_{max} for Model I and Model II



It must be noted that although Model II has a shorter model run time and more optimal solutions, most numerical examples in both Models show good results with short run times and with 0% relative gaps. For these problem sizes, the models can generate good results in a reasonable time and since the differences between both Models are not considerable, Model I might propose a better set of solution for decision makers since it minimizes the total workload difference amongst rooms instead of only minimizing the maximum workload difference between any two rooms.

Chapter VI - CONCLUSIONS

This chapter concludes this dissertation. The main results and the limitations on the work are emphasized, and directions for future research on this topic are recommended.

6.1 Concluding Remarks

Generally speaking, scheduling systems in practice have a lot of variables that are not fully optimized. For instance, take elective patients: they are treated according to a FCFS (First Come First Serve) approach on the waiting lists, without considering the evolution of the patients' health status, patient characteristics and patient categorization. Additionally, regarding the resources available within the clinics or hospitals, most of the times the process is not efficient, as for instance, room availability can be irregular and doctor allocation and their respective workloads can be unfair. Moreover, research on prioritization procedures and resource balance is very scarce and is done separately from the scheduling procedures. These problems lead to workload imbalance and workload imbalance leads to inefficient appointment schedules.

In this dissertation, a two-model approach is proposed for designing a scheduling system in an outpatient specialty clinic or hospital, providing services of multiple types and with room availability by said specialty. These service types are distinguished by time measures, being categorized by one greater than the other. To this end, an example to consider could be patient heterogeneity. Take for instance the case where there are two types of patients. This categorization could be first-time patients and returning patients. It is common to divide patients into newly referred ('new') patients and follow up ('return') patients. New patients often require a longer consultation time compared to returning patients (Deceuninck et al., 2018).

As result, the proposed two-model approach is tested for scenarios with different levels of patient demand for said service types, with different number of rooms available and different number of specialties. In the first Model approach, an Integer Linear Programming model is formulated, known as Model I, with the objective of minimizing workload differences among rooms within the outpatient clinic. In the second Model approach, the Integer Linear Programming model known as Model I is reformulated with a new objective function, that proposes the minimization of the maximum workload difference, with a $minxmax$ decision process.

The computational experiments show that the model returns good results concerning computational efficiency (most gaps are 0%) in a short period of time (1-hour computation time limit). Results obtained demonstrate that the proposed two-model approach can efficiently identify promising scheduling designs for an outpatient clinic on an average

personal computer. The best daily demand assignment found can improve the balance within workloads, being on Model I and Model II.

For outpatient appointment demand assignment, it is likely to achieve higher-quality solutions when integrating the first approach, i.e., Model I, since although results are fairly similar for both Models, the objective of minimizing workloads is better observed when the focus is on minimizing the total differences of workloads between rooms and not only the maximum value of workload difference for any room.

Meanwhile, results suggest that the sampling based solutions to both Models become more sensitive as the demand and the number of specialties increase and with their respective number of rooms available.

The results obtained with the computational experiments allow answering the research question: "Is it possible to minimize workload differences given a demand for a clinical specialty and a service type and their respective allocation to available rooms, optimizing the clinics' available resources based on a number of appointments that should be scheduled daily?" by providing an optimal number of appointments reserved for each specialty and each service type within each room. The results obtained suggest that it is indeed possible. Although not the main focus of analysis, the model output is the result with foremost importance to decision makers, and it can easily be adapted to different scenarios and contexts.

6.2 Limitations and future research

To conclude this dissertation, some limitations of work are enumerated, which may be used to improve the proposed approach and explore future work and research directions.

The computational experiments show that the implemented mathematical model returns results in a reasonably short period of time (1-hour computation limit is used). Nevertheless, although the models are reliable, they do not consider some aspects of real scenarios in hospitals such as the uncertainty of the duration of the appointments, cancellations, walk-ins, no-shows and other environmental factors as described in the dissertations' literature review.

The two-model approach proposed in this dissertation provides a quantitative tool for outpatient specialty clinics to design better appointment scheduling systems. These daily demand assignment could be adapted as weekly or monthly schedules.

In addition, although generated numerical examples are built to be realistic, it would be important to test and validate these approaches in outpatient clinics with real data. Namely, it is of interest to understand the impact on the performance of the models regarding demand, waiting lists and dimensions of the problems to the results obtained.

Furthermore, adding even more service types and categorizations to the parameters is an important future approach as it makes the models more realistic and viable to problems with increased dimensions.

The approach of this dissertation is generalized, as it names no specific specialty and service type. Adaptation to real clinics and hospitals in case studies is an important factor in order to validate the models. In future researches, the focus should be on investigating whether such an integrated model could find schedules that significantly improve the performance metrics.

Moreover, these approaches can be extended to other queuing systems. Indeed, the programming models can be applied in several contexts, according to the parameters set. The model can be adapted to appointments, meetings or even queues in customer service for any areas and be used in their scheduling process.

REFERENCES

Ahmadi Javid, A., Jalali, Z., & Klassen, K. (2016). Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, *258*. doi: 10.1016/j.ejor.2016.06.064

Babes, M., & Sarma, G. V. (1991). Out-patient queues at the ibn-rochd health centre. *Journal of the Operational Research Society*, *42*(10), 845-855. doi: 10.1057/jors .1991.165

Bailey, N. T. J. (1952). 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.

Balasubramanian, H., Biehl, S., Dai, L., & Muriel, A. (2013). Dynamic allocation of same-day requests in multi-physician primary care practices in the presence of prescheduled appointments. *Health care management science*, *17*. doi: 10.1007/s10729-013-9242-2

Barghash, M., & Saleet, H. (2018). Enhancing outpatient appointment scheduling system performance when patient no-show percent and lateness rate are high. *International Journal of Health Care Quality Assurance*, *31*, 00-00. doi: 10.1108/IJHCQA-06 -2015-0072

Buhaug, H. (2002). Long waiting lists in hospitals. *BMJ*, *324*(7332), 252–253. doi: 10.1136/bmj.324.7332.252

Cayirli, T., & Gunes, E. D. (2014). Outpatient appointment scheduling in presence of seasonal walk-ins. *Journal of the Operational Research Society*, *65*(4), 512-531. doi: 10.1057/jors.2013.56

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, *12*, 519 - 549. doi: 10.1111/ j.1937-5956.2003.tb00218.x

Cayirli, T., Veral, E., & Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health care management science*, *9*, 47-58. doi: 10.1007/s10729-006-6279-5

Cayirli, T., Veral, E., & Rosen, H. (2008). Assessment of patient classification in appointment system design. *Production and Operations Management*, *17*(3), 338-353. doi: https://doi.org/10.3401/poms.1080.0031

Cayirli, T., & Yang, K. K. (2014). A universal appointment rule with patient classification for service times, no-shows, and walk-ins. *Service Science*, *6*(4), 274-295. doi: 10.1287/serv.2014.0087

Cayirli, T., Yang, K. K., & Quek, S. A. (2012). A universal appointment rule in the presence of no-shows and walk-ins. *Production and Operations Management*, *21*(4), 682-697. doi: https://doi.org/10.1111/j.1937-5956.2011.01297.x

Conforti, D., Guerriero, F., & Guido, R. (2008). Optimization models for radiotherapy

patient scheduling. *4OR*, *6*, 263-278. doi: 10.1007/s10288-007-0050-8

Conforti, D., Guerriero, F., & Guido, R. (2010). Non-block scheduling with priority for radiotherapy treatments. *European Journal of Operational Research*, *201*(1), 289-296. doi: https://doi.org/10.1016/j.ejor.2009.02.016

Conforti, D., Guerriero, F., & Guido, R. (2011). An optimal decision-making approach for the management of radiotherapy patients. *OR Spectr.*, *33*(1), 123–148. doi: 10.1007/s00291-009-0170-y

Cox, T. F., Birchall, J. P., & Wong, H. (1985). Optimising the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics*, *12*(2), 113-126. doi: 10.1080/02664768500000017

Dantas, L. F., Fleck, J. L., Cyrino Oliveira, F. L., & Hamacher, S. (2018). No-shows in appointment scheduling – a systematic literature review. *Health Policy*, *122*(4), 412-421. doi: https://doi.org/10.1016/j.healthpol.2018.02.002

Deceuninck, M., Fiems, D., & De Vuyst, S. (2018). Outpatient scheduling with unpunctual patients and no-shows. *European Journal of Operational Research*, *265*(1), 195-207. doi: https://doi.org/10.1016/j.ejor.2017.07.006

Denton, B., & Gupta, D. (2003). A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, *35*, 1003-1016. doi: 10.1080/07408170304395

Drupsteen, J., van der Vaart, T., & van Donk, D. (2013). Integrative practices in hospitals and their impact on patient flow. *International Journal of Operations and Production Management*, *33*(7), 912–933. doi: 10.1108/IJOPM-12-2011-0487

Erdogan, S. A., & Denton, B. (2013). Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing*, *25*(1), 116-132. doi: 10.1287/ijoc.1110.0482

Gallucci, G., Swartz, W., & Hackerman, F. (2005). Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*, *56*(3), 344-346. doi: 10.1176/appi.ps.56.3.344

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*, 800-819. doi: 10.1080/07408170802165880

Gupta, D., & Wang, L. (2008). Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, *56*(3), 576-592. doi: 10.1287/opre.1080.0542

Heaney, D. J., Howie, J. G., & Porter, A. M. (1991). Factors influencing waiting times and consultation times in general practice. *British Journal of General Practice*, *41*(349), 315–319.

Ho, C.-J., & Lau, H.-S. (1992). Minimizing total cost in scheduling outpatient appointments. *Management Science*, *38*, 1750-1764. doi: 10.1287/mnsc.38.12.1750

Hulshof, P. J. H., Kortbeek, N., Boucherie, R. J., Hans, E. W., & Bakker, P. J. M. (2012). Taxonomic classification of planning decisions in health care: a structured review of the state of the art in or/ms. *Health Systems*, *1*(2), 129-175. doi: 10.1057/

hs.2012.18

Kaandorp, G., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health care management science*, *10*, 217-29. doi: 10.1007/s10729-007-9015-x

Klassen, K., & Yoogalingam, R. (2013). Appointment system design with interruptions and physician lateness. *International Journal of Operations and Production Management*, *33*. doi: 10.1108/01443571311307253

Klassen, K., & Yoogalingam, R. (2019). Appointment scheduling in multi-stage outpatient clinics. *Health Care Management Science*, *22*. doi: 10.1007/s10729-018-9434-x

Klassen, K. J., & Rohleder, T. (2004). Outpatient appointment scheduling with urgent clients in a dynamic, multi period environment. *International Journal of Service Industry Management*, *15*, 167-186. doi: 10.1108/09564230410532493

Klassen, K. J., & Yoogalingam, R. (2009). Improving performance in outpatient appointment services with a simulation optimization approach. *Production and Operations Management*, *18*(4), 447-458. doi: https://doi.org/10.1111/j.1937-5956.2009.01021.x

Klassen, K. J., & Yoogalingam, R. (2014). Strategies for appointment policy design with patient unpunctuality. *Decision Sciences*, *45*(5), 881-911. doi: https://doi.org/10.1111/deci.12091

Koeleman, P. M., & Koole, G. M. (2012). Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Transactions on Healthcare Systems Engineering*, *2*(1), 14-30. doi: 10.1080/19488300.2012.665154

Krishnamoorthy, A., Pramod, P., & Chakravarthy, S. (2014). Queues with interruptions: a survey. *TOP: An Official Journal of the Spanish Society of Statistics and Operations Research*, *22*(1), 290-320.

Kuiper, A., Kemper, B., & Mandjes, M. (2015). A computational approach to optimized appointment scheduling. *Queueing Systems*, *79*(1), 5–36. doi: 10.1007/s11134-014-9398-6

Lehaney, B., Clarke, S. A., & Paul, R. J. (1999). A case of an intervention in an outpatients department. *Journal of the Operational Research Society*, *50*(9), 877-891. doi: 10.1057/palgrave.jors.2600796

Liu, L., & Liu, X. (1998a). Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, *49*(12), 1254-1259. doi: 10.1057/palgrave.jors.2600631

Liu, L., & Liu, X. (1998b). Block appointment systems for outpatient clinics with multiple doctors. *Journal of the Operational Research Society*, *49*(12), 1254-1259. doi: 10.1057/palgrave.jors.2600631

Liu, L., & Liu, X. (1998c). Dynamic and static job allocation for multi-server systems. *IIE Transactions*, *30*(9), 845-854. doi: 10.1080/07408179808966530

Liu, N., Finkelstein, S., Kruk, M., & Rosenthal, D. (2017). When waiting to see a

doctor is less irritating: Understanding patient preferences and choice behavior in appointment scheduling. *Management Science*, *64*. doi: 10.1287/mnsc.2016.2704

Marynissen, J., & Demeulemeester, E. (2019). Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, *272*, 407-419. doi: 10.1016/j.ejor.2018.03.001

Mayer, G., Villaire, M., & Connell, J. (2005). Ten recommendations for reducing unnecessary emergency department visits. *The Journal of nursing administration*, *35*, 428-30. doi: 10.1097/00005110-200510000-00003

Meza, J. (1998). Patient waiting times in a physician's office. *The American journal of managed care*, *4*(5), 703—712.

Moore, C., Wilson-Witherspoon, P., & Probst, J. (2001). Time and money: effects of no-shows at a family practice residency clinic. *Family medicine*, *33*(7), 522—527.

Murray, M., & Tantau, C. (2000). Same-day appointments: exploding the access paradigm. *Family practice management*, *7*(8), 45—50.

Muthuraman, K., & Lawley, M. (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, *40*(9), 820-837. doi: 10.1080/07408170802165823

Oppenheim, G. L., Bergman, J., & English, E. C. (1979). Failed appointments: a review. *Journal of Family Practice*, *8*, 789-796.

Pesata, V., Pallija, G., & Webb, A. (1999). A descriptive study of missed appointments: families' perceptions of barriers to care. *Journal of pediatric health care : official publication of National Association of Pediatric Nurse Associates; Practitioners*, *13*(4), 178—182. doi: 10.1016/s0891-5245(99)90037-8

Publico. (2017). Mais de 75% dos médicos admitem trocar o sns pelo sector privado. *Alexandra Campos*. Retrieved from `https://www.publico.pt/2017/01/12/sociedade/noticia/mais-de-75-dos-medicos-admite-trocar-o-sns-pelo-sector-privado-1758095`

Pérez, E., Ntaimo, L., Wilhelm, W. E., Bailey, C., & McCormack, P. (2011). Patient and resource scheduling of multi-step medical procedures in nuclear medicine. *IIE Transactions on Healthcare Systems Engineering*, *1*(3), 168-184. doi: 10.1080/19488300.2011.617718

Qu, X., Peng, Y., Kong, N., & Shi, J. (2013). A two-phase approach to scheduling multi-category outpatient appointments – a case study of a women's clinic. *Health care management science*, *16*. doi: 10.1007/s10729-013-9223-5

Qu, X., Peng, Y., Shi, J., & LaGanga, L. (2015). An mdp model for walk-in patient admission management in primary care clinics. *International Journal of Production Economics*, *168*, 303-320. doi: https://doi.org/10.1016/j.ijpe.2015.06.022

Qu, X., Rardin, R. L., Williams, J. A. S., & Willis, D. R. (2007). Matching daily healthcare provider capacity to demand in advanced access scheduling systems. *European Journal of Operational Research*, *183*(2), 812 - 826. doi: https://doi.org/10.1016/

j.ejor.2006.10.003

Rising, E. J., Baron, R., & Averill, B. (1973). A systems analysis of a university-health-service outpatient clinic. *Operations Research*, *21*(5), 1030-1047. doi: 10.1287/opre.21.5.1030

Robinson, L., & Chen, R. (2003). Scheduling doctor's appointments: Optimal and empirically-based heuristic policies. *Iie Transactions*, *35*, 295-307. doi: 10.1080/07408170304367

Robinson, L. W., & Chen, R. R. (2010). A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing and Service Operations Management*, *12*(2), 330-346. doi: 10.1287/msom.1090.0270

Samorani, M., & Ganguly, S. (2016). Optimal sequencing of unpunctual patients in high-service-level clinics. *Production and Operations Management*, *25*(2), 330-346. doi: https://doi.org/10.1111/poms.12426

Samorani, M., & Harris, S. (2019). The impact of probabilistic classifiers on appointment scheduling with no-shows. *SSRN Electronic Journal*. doi: 10.2139/ssrn.3463501

Samorani, M., & LaGanga, L. R. (2015). Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research*, *240*(1), 245 - 257. doi: 10.1016/j.ejor.2014.06.034

Soriano, A. (1966). Comparison of two scheduling systems. *Operations Research*, *14*(3), 388-397. doi: 10.1287/opre.14.3.388

Tai, G., & Williams, P. (2012). Optimization of scheduling patient appointments in clinics using a novel modelling technique of patient arrival. *Computer Methods and Programs in Biomedicine*, *108*(2), 467 - 476. doi: https://doi.org/10.1016/j.cmpb.2011.02.010

Tang, J., Yan, C., & Fung, R. Y. (2014). Optimal appointment scheduling with no-shows and exponential service time considering overtime work. *Journal of Management Analytics*, *1*(2), 99-129. doi: 10.1080/23270012.2014.915127

Turkcan, A., Zeng, B., & Lawley, M. (2012). Chemotherapy operations planning and scheduling. *IIE Transactions on Healthcare Systems Engineering*, *2*(1), 31-49. doi: 10.1080/19488300.2012.665155

Van de Vrugt, Noelle Maria. (2016). *Efficient healthcare logistics with a human touch* (Doctoral dissertation, University of Twente, Netherlands). doi: 10.3990/1.9789036541152

Vanberkel, P., Boucherie, R., Hans, E., Hurink, J., & Litvak, N. (2010). A survey of health care models that encompass multiple departments. *International journal of health management and information (IJHMI)*, *1*(1), 37-69.

Vissers, J. (1979). Selecting a suitable appointment system in an outpatient setting. *Medical care*, *17*(12), 1207–1220.

Walter, S. (1973). A comparison of appointment schedules in a hospital radiology department. *British Journal of Preventive and Social Medicine*, *27*(3), 160—167. doi:

10.1136/jech.27.3.160

Wang, P. P. (1993). Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics (NRL)*, *40*(3), 345-360. doi: https://doi.org/10.1002/1520-6750(199304)40:3<345::AID-NAV3220400305>3.0.CO;2-N

Wang, W. Y., & Gupta, D. (2011). Adaptive appointment systems with patient preferences. *Manufacturing and Service Operations Management*, *13*(3), 373-389. doi: 10.1287/msom.1110.0332

White, D. L., Froehle, C. M., & Klassen, K. J. (2011). The effect of integrated scheduling and capacity policies on clinical efficiency. *Production and Operations Management*, *20*(3), 442-455. doi: https://doi.org/10.1111/j.1937-5956.2011.01220.x

White, M. J. B., & Pike, M. C. (1964). Appointment systems in out-patients' clinics and the effect of patients' unpunctuality. *Medical Care*, *2*(3), 133–145.

Wijewickrama, A., & Takakuwa, S. (2008). Outpatient appointment scheduling in a multi facility system. In *Proceedings of the 40th conference on winter simulation* (p. 1563–1571). Winter Simulation Conference.

Yang, K. K., Lau, M. L., & Quek, S. A. (1998). A new appointment rule for a single-server, multiple-customer service system. *Naval Research Logistics (NRL)*, *45*(3), 313-326. doi: https://doi.org/10.1002/(SICI)1520-6750(199804)45:3<313::AID-NAV5>3.0.CO;2-A

Zacharias, C., & Pinedo, M. (2014). Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, *23*(5), 788-801. doi: https://doi.org/10.1111/poms.12065

Zhu, H., Chen, F., Leung, E., & Liu, X. (2017). Outpatient appointment scheduling with unpunctual patients. *International Journal of Production Research*, 1-21. doi: 10.1080/00207543.2017.1355574

UNIVERSIDADE DOS AÇORES
Faculdade de Economia e Gestão

Rua da Mãe de Deus
9500-321 Ponta Delgada
Açores, Portugal