A Framework for online Multiple Examinations Scheduling Problem in speciality clinic

Sijia Yang

Supervisor: Prof Philip Treleaven

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Msc Data Science and Machine Learning
of the
University College London

Department of Computer Science
University College London

September 7, 2018

I,Sijia Yang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis investigates how to effectively schedule multiple examinations in the prostate cancer pathway under different hospital environments. This work is important as an effective scheduling for examinations is crucial to patients in the cancer pathway and it may be largely affected by the uncertainties, such as crowdedness and urgency level, in the complex hospital environment. Therefore, a comprehensive software framework, called SlotRkr, is proposed as a solution to address the multiple examinations scheduling problem. It comprises two components that provide scheduling and simulating functionalities. The experiments in this thesis investigate two main challenges faced by most specialty clinics when time-based performance metrics are considered:

1. The most efficient scheduling method with optimal performance that applied in current hospital environment;
2. The most robust scheduling method that can dynamically adapted to changing environment.

Scheduling Model

The first component in SlotRkr is a scheduling model aiming at creating effective schedule plan upon patient's arrival for One-Stop-Clinic in The Princess Alexandra Hospital (PAH). In the complex clinical environment, the interactions between patients, multiple examinations providers and scheduler are firstly modelled by multi-agent approach. Then, a heuristic method is developed to provide an optimal solution based on the future information obtained in advanced. Although this method is not realistic in practice, it provides idea to develop a novel cost-based scheduling method that dynamically adapts to the changing environment. For comparison, the currently used scheduling method, first-come-first-serve, is also included in this model.

Simulation Model

The second component in SlotRkr constructs mathematical simulation model to simulate the patients' arrival streams to the One-Stop-Clinic. The data provided by The Princess Alexandra Hospital (PAH) is first re-sampled using bootstrapping technique and further

classified according to patients' status. With a 95% confidence level of Chi-Square Goodness-Of-Fit test, patients' arrival process can be modelled by Poisson distribution and the probability of patients' random status is modelled by beta distribution. By passing different parameter to these distributions, the simulation model can be configured to simulate different crowdedness and urgency level in the hospital environment.

Experiments

A series of theoretical experiments are carried out to analyse how SlotRkr tackles the main challenges stated before. There are in total 21 experimental units in the experiments. In each unit, different configuration is passed to the simulation model to generate patient streams and the heuristic, cost-based and first-come-first-serve scheduling method are applied to schedule such streams. The results of experiments are examined with respects to three hypotheses and it shows that first-come-first-serve is suggested to apply in current hospital environment while cost-based scheduling method is more robust as the hospital environment changes.

Contributions to Science:

The major contribution of this thesis is to develop a comprehensive software framework to address the multiple examinations scheduling problem. The framework consists of

i. *a computer-based simulation model* that captures the randomness of patients' arrival process and their corresponding categories;

ii. *a multi-agent system* that models the patients and departments in hospitals and create effective schedule plans by the novel method SlotRkr;

The framework firstly contributes to the existing literature in the way that it models the hospitals environment. The number of patients per week is measured by "crowdedness" level in experimental settings and controlled by parameter in a Poisson distribution; the average urgency level of patients' scheduling requests is controlled by the parameter of the beta distribution for a certain group. In this way, the simulation process is interpretable and repeatable.

Furthermore, the novel cost-based scheduling method is intuitive, robust and easy to implement. It considers different performance metrics in the cost function and assigns different weights to them to represent a trade-off between them, which can be further adjusted by users for to address different problem. I also introduce a penalty component in the cost function to penalise those schedules that assign early timeslots to less urgent patients. It is similar to humans' consideration when making schedule while it is seldom modelled in previous literature.

Acknowledgement

I would like to thank my supervisor Prof. Philip Treleaven and his PhD student Almasi Sirvan who offer the opportunity of my MSc project. My most sincere gratitude goes to Prof. Philip Treleaven for his continuous help with my thesis. Also, my deepest thanks go to Almasi Sirvan due to his patient guidance and valuable advice on my project.

I would like to show gratitude to the members in The Princess Alexandra Hospital (PAH), Anne and Jonathan, who explain the details related to the prostate cancer pathway as well as the provision of data used in this thesis.

I would also like to thank my parents for their unconditional love and continuous support during this time spent literally an ocean apart from them.

Finally, a big thank you goes to my good friend and room-mate who always accompanies me during the period I was doing my project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter presents an overview of the thesis. First, the background of problem investigates in this thesis is briefly introduced. In section 1.2, the motivation and the main challenges of this research are explained. Furthermore, the objectives are claimed in section 1.3. Finally, the structure of the whole thesis is given in section 1.4.

1.1 Problem Background

Prostate cancer is the most common cancer affecting men in United Kingdom, with over 40,000 new cases diagnosed every year [18]. According to [17], it is causing more death due to its stable mortality rate and increasing incidence rate. The prevalence of prostate cancer is not only affecting patients' life, but also has a negative social impact regrading to huge medical expenditure and possible pressure given by family members and friends [62]. In order to mitigate these undesired effects, nationally agreed and clinically endorsed pathway is accepted in prostate cancer system.

Clinical pathways (or critical pathways) are structured multidisciplinary plans that describe the sequences and timing of activities that hospital staff provide to patients [43]. The whole prostate cancer pathway can be divided into three phases: the diagnostic phase, the treatment phase and the support phase. Diagnostic pathways illustrate how timely and effective care can be provided to patients presenting with cancer symptoms. Treatment pathway is a tailored treatment plan for every men with detailed information about the treatments available for the stage of their disease, the side effects and the outcomes of each treatments. Support pathway aims at delivering good supportive care that ‘better understand the needs of those living with cancer and develop models of care that meet their needs” [49].

In this thesis, a diagnostic prostate cancer pathway is studied in a speciality clinic, called “One-Stop-Clinic” (OSC) in The Princess Alexandra Hospital (PAH). One of the essential components of pathway is the time-line, which consists of possible activities during a pathway along with the corresponding due dates. As shown in Fig.1.1, there are five activities included in the diagnostic pathway. The official due date set by NHS for each activity is called “Maximal acceptance” due date, which is demonstrated as the time-line at the top of the figure, while the “Good practice” time-line provided by PAH is shorter, which is shown at the bottom of the figure. Each activity is briefly described below.

Activity 1: Urgent GP Referral

Before patients being referral to the local hospital, they will receive a pre-test on GP. According to the pre- test results, they are further divided into two types: urgent patients and regular patients. Only urgent referrals are accepted by OSC while regular patients who have lower risk of having prostate cancer are referred to other clinic. When referrals are accepted by OSC, they will enter the prostate cancer pathway and this date is indicated as ”DAY0”.

Activity 2: Clinical assessment

OSC provides ”Fast Track (F/T)” slots for urgent referrals to have their first appointment with urologists, which is called “clinical assessment”. The assessment lasts about 20 minutes and patients may get examinations referral afterwards. Patients are encourage to visit the OSC within 7 days after they got GP referral.

Activity 3: MRI scan

After clinical assessments, most patients should take two examinations sequentially, which help to detect the type and size of tumour. Magnetic resonance imaging (MRI) scan is the first examination and it is suggested to happen at the same day or in the next day. Report of such examination will be released at the same day.

Activity 4: TRUS biopsy test

With the MRI scan report, patients are allowed to take Transrectal ultasound (TRUS)

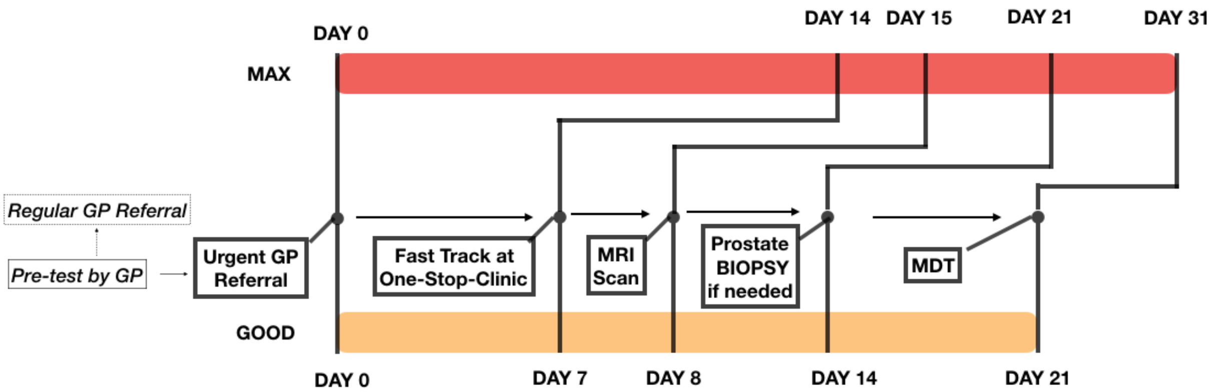

Figure 1.1: Pathway time-line

biopsy test. Ac- cording to maximal acceptance standard, it should happen by "DAY21" in the pathway, while PAH's good practice pursue to shorten it to 14 days because this test has a longer period (3 to 7 days) for the report releasing, which may affect the later activities.

Activity 5: MDT

When two examinations finish, decisions about patients will be dis- cussed in an Multi-disciplinary Team (MDT) meeting. It is the mark of the end of diagnostic phase. The total length of diagnosis phase is defined as the days between the referral date and MDT date, which is suggested to be less than 31 days (maximal acceptance) or 21 days (good practice).

1.2 Motivation

The main issue with current prostate cancer pathway is the inefficient delivery of some intermediate activities, resulting in a deviation from original pathway [16]. As discussed above, the activities before diagnosis include consultation, MRI scan and biopsies. Since consultation date might be related to patients' characteristics and external factors such as weather and unanticipated events, which cannot be controlled by OSC, the possible solution to mitigate this problem is to effectively schedule patients to the two examinations in a way that most patients will not miss their diagnosis deadline. In this case, it is natural to classified patients into groups with different urgency levels according to their arriving days as the earlier they arrive for consultation, the more time will be left for them to complete the rest examinations. Therefore, this problem can be treated as a multi-priority multi-appointment scheduling problem for outpatients.

Operation Research (OR) has enjoyed a long tradition of addressing logistics healthcare challenges, especially resource capacity planning and appointment scheduling issues [6]. Furthermore, the advanced study on variational inequalities and dynamic programming allows OR technique to solve dynamic problems. One successful example of applying OR technique to solve a multi-priority patient-scheduling problem can be review in [47]. With the aim of reducing the accessing time to diagnosis, their model proved to be outstanding regardless of the setting of clinics and size of hospitals. Nevertheless, lots of research adopting methods such as Markov Decision Process (MDP) and Integer Linear Programming (ILP) aim at solving the mismatch between static resource and dynamic patient flow each day and they usually make schedule at the end of each day based on the collected information during that day.

Even though the problem investigated in this thesis shares some common characteristics with outpatient scheduling problem, such as a stochastic patient arrivals and different levels of urgency of patients, there are some key differences:

- Patients need to have the date and time for later examination right after the consultation slots, resulting in a real-time scheduling system, which means it is not possible to wait for the accumulation of patients before generating a schedule;
- Outpatient scheduling problems usually consist of scheduling a single doctor appointment for a patient, which often has a stochastic duration. When follow-up sessions exist, they may be scheduled on the fly [25]. However, prostate examinations scheduling consists of booking several time slots of deterministic duration such that consecutive session of the same patient are scheduled a pre-determined number of days apart;
- There is a fixed sequences of examinations, which means patients need to finish the

previous appointment before they can attend the next appointment. Thus, the local performance of each examination should be considered.

These features make the problem not only hard to solve, but also hard to model. For example, even in a simple scenario, solving a MDP or ILP requires one hour [13], which is not suitable for the case in OSC where each scheduling decision should be made in a very short time. Additionally, optimizing local performance is often complex. A department receives appointment requests from outpatient departments with varying medical properties and urgencies. With a fixed resource capacity, appointments must be scheduled such that for all urgency levels a satisfactory fraction of patients is scheduled on time. The typical approach to this problem is to allocate parts of the resource capacity to each patient group. Allocating capacity specifically can indeed improve performance depending on problem properties. However, due to fluctuations in patient arrivals, initial capacity allocation can regularly mismatch current demands [57].

However, it is worthwhile to investigate a real-time scheduling approach to support the examinations scheduling as it provides a series of benefits:

- **Increased Efficiency** The reason why OSC need real-time appointment schedules is that there are a large number of examination slots and consultation slots available on the same day, as the aim of OSC is to efficiently provide patients with all needed care services on one day.
- **Reduced anxiety** OSC found out that patients with suspicious cancer are usually anxiety about their future stages in hospital when they finish the consultation. If they can know a blueprint that contains the exact date and time of all examinations they need to take immediately, they may be less worried.
- **Decreased pathway length** With an effective scheduling system, the probability of scheduling patients to the date that exceeds their diagnosis deadline will be decreased, resulting in a shorter pathway length for diagnosis phase, which means a faster access to treatment and it is significant to cancer patients.

1.3 Objective

The research presented in this thesis focuses on developing and evaluating models and algorithms used to automatically provide suspected prostate cancer patients with schedule plan that consists of schedules for multiple examinations in prostate cancer pathway. The main objectives of this thesis are:

By developing a comprehensive framework, patients random arrival process can be modelled and effective schedule plan that consists of two schedule dates for two examinations in the prostate cancer pathway can be create dynamically upon each patient's arrival such that the scheduling efficiency in the clinic can be improved in terms of the ratio of patients who exceed their due date of finishing their pathway.

To achieve this objective, there are four main tasks for this research:

1. Review literatures that solve patient scheduling problem by different methodologies, such as computer simulation, heuristics, Markov processes, mathematical programming and queueing theory, so as to identify the advantage and limitations of each method and select the methodology used in this thesis.

2. Analyse the data provided by PAH to identify hidden patterns of patients' arrival and their arriving status so as to construct mathematical model to simulate patients' arrival process.

3. Develop robust methods capable of creating a real-time schedule in an acceptable amount of time.

4. Analyse the performance of the developed methods by a detailed and well structured set of experiments.

1.4 Structure of this Thesis

Structure of this thesis is as follows,

- Chapter 2 examines different categories of patient scheduling problems. The majority of the literature relevant for this work is discussed. Papers are classified according to the methods used.

- Chapter 3 details the problem that currently faced by PAH and proposes the formal solution to that problem. The experiments are carried out to test the solution and the experimental result analysis is also given.

- Chapter 4 summaries this research and outlines possibilities for future work.

Chapter 2

Background and Literature Review

This chapter aims to deliver some background information of the prostate cancer system and common techniques used in appointment scheduling (AS) system. In order to better understand the available literature, a classification scheme based on the works by [21] is presented. Section 2.1 describes a classification according to the environment the problem deals with. Section 2.2 lists the most common characteristics in outpatient scheduling and how they are more commonly found in each environment. The most common performance measures are described in section 2.3 and common methods used to tackle outpatient scheduling problems are introduced in section 2.4.

2.1 Outpatient Scheduling Environments

Outpatient scheduling systems can be classified according to the environment for which they are designed. Three main types of scheduling environments that defined in [21] is briefly introduced below while the problem focused in this thesis belongs to the second category:

Primary care: Patients arrive to a clinic or hospital, usually to see a general practitioner (GP).In the majority of environments, patients call in advance to schedule an appointment. The scheduler then books a time and day for the patient's appointment. There may also be walk-in patients, who must be accommodated sometime during the day. Some of these may be emergency or urgent patients and may need to be seen immediately.

Specialty care: Patients are usually referred from primary care to a specialty clinic in order to receive treatment specific to the patient's diagnoses. Specialty clinics focus on specific, often complex, diagnoses and treatments. For the majority of specialty clinics, patients must be referred to them by a GP before they can ask for an appointment. Session durations are usually deterministic in this environment. They may also have a large variation due to different diagnoses and, therefore, fixed length slots are not usually adopted. Instead, each session is booked to use only the amount of time it requires.

Resources in specialty clinics are often very expensive, either in the form of equipment and machinery, or in the form of specialised doctors. For this reason, achieving high levels of utilisation in these environments is highly desired in order not to waste resource time. Patients commonly have different levels of urgency, where normal patients can be scheduled well in advance and emergency or urgent patients can arrive with short notice and need to be treated immediately. The main challenge in this environment usually resides in reserving enough capacity for patients of high urgency, while maintaining a high utilisation of the resources.

Elective surgeries: Scheduling of chosen and planned in advance surgery procedures in operating rooms. The biggest difference between surgery scheduling and other scheduling is that the duration of surgeries is not known a priori. Complications can occur during surgery, which can greatly increase their duration. This stochastic duration with large variation makes surgery scheduling one of the most complex patient scheduling problems. [8] and [39] present reviews of scheduling algorithms tailored for this environment. Also, surgeries require a greater variety of resources to be allocated to them when compared to the primary and specialty care environments.

2.2 Factors in Outpatient Scheduling

In this section, five relevant factors that are encountered in appointment scheduling environments are introduced. By identify these key factors, the modelling of the scheduling problem in this thesis is briefly explained as well.

Operational planning involves the short-term decision making related to the execution of the healthcare delivery process. There are two types of planning: "offline" and "online" which reflect the nature of a static and dynamic healthcare process, respectively.

i. *offline* operational planning concerns the in advance planning of operations. It comprises the detailed coordination of the activities regarding current (elective) demand. Examples of offline operational planning are: treatment selection, appointment scheduling, nurse rostering, inventory replenishment ordering, and billing [25].

ii. *online* operational planning involves control mechanisms that deal with monitoring the process and reacting to unforeseen or unanticipated events. Examples of online planning functions are: triaging, add-on scheduling of emergencies, replenishing depleted inventories, rush ordering surgery instrument sterilization, handling billing complications [25]. Online literature includes patient scheduling [37] and (open) shop scheduling problems [1]. The more closely related problem of appointment scheduling and real-time capacity allocation in an MRI setting was addressed by [20]. The use of approximate dynamic programming to solve the problem of dynamically allocating diagnostic imaging resources to multiple patient priority classes, in order to achieve targeted wait times was investigated in [47].

Number of Services determines the number of service that patients are waiting for. Almost all studies in the literature model a single-stage system where patients queue for a single service. However, in our problem, patients are waiting for multiple examinations. Thus, a multiple stage system is considered. A few simulation studies investigate clinic environments where a patient may pass through facilities such as registration, pre-examination, post-examination, x-ray, laboratory, checkout, etc. [12, 50, 54]. In such multi-stage models, the patient flow (transition) probabilities associated with each facility need to be specified for Markov Process modeling.

Scheduling combination appointments involves complex local scheduling. Work such as [47] discuss such problems. Also, in [20] the authors discuss a local profit maximization problem of a MRI scheduling problem for three classes of patients. Their more abstract model requires setting specific revenue and penalty functions, for which the authors identify properties of an optimal solutions. The authors focus on local performance and do not consider a trade-off against patient preferences.

A similar research is conducted in [58], where a new multi-agent Pareto-improvement appointment exchanging algorithm is developed. This algorithm starts from a simple schedule such as first-come-first-serve on several diagnostic resources and interchanges the existing appointments that have been scheduled using single-resource algorithms by using different agents (patient agents, resource agents, staff agents) to improve the existing schedules [40].

The Arrival Process describes several arrival characteristics of patients. According to [9], it consists of the following factors, which affect appointment system performance. Each factor and related literature is described below to introduce the problem setting of this thesis.

i. *Unpunctuality of patients* can be defined as the difference between a patient's appointment time and actual arrival time. Lots of research focus on the same day unpunctuality. For example, [3, 7, 19, 32, 36, 60, 61] conduct research and empirical evidence suggests that patients arrive early more often than late. However, in this problem, the same day unpunctuality is not considered due to the specialty of OSC clinic. It receive outpatients' consultation appointment only on Tuesday and the unpunctuality remains in the weekly demand. It can also be treated as "no-shows" in current week, which is described below. Some authors use independent random variable with a certain limit on maximum lateness to model patients' lateness [42] while others use fitted theoretical probability distributions to empirically derived histograms of patient arrival times relative to their appointment to model that [12, 19, 53]. In either ways, it is assumed that patients' unpunctuality is independent of their scheduled appointment times.

ii. *Presence of no-shows* is moderately studied in the literature using no-show probabilities (p) that range from 5 to 30 percent. It is considered to affects the performance and the choice of an AS in lots of research [23]. Some focus on the study of the possible variables (such as age, socioeconomic level, etc.) that might affect this patient attendance [15] and some focus on finding the optimal number of patients to overbook in a day to reduce the impact of no-shows [31,33,34,44]. In this problem, however, no-shows is not included in the environment factors. The reason is the the appointment book by GP has no specific days, means that patients can come and visit the clinic whenever they have the appointment [55].

iii. *Presence of walk-ins (regular and emergency)* is neglected in most studies as well as in this thesis. In the U.K., hospital clinics are primarily used for consultation services for patients referred to them by the general practitioner outside the hospital, and walk-ins are rarely accepted [9].

Service Times can be defined as the sum of all the times a patient is claiming the doctor's attention, preventing him/her from seeing other patients [4]. The majority of the studies assume patients are homogeneous for scheduling purposes, and use independently and identically distributed (i.i.d.) service times for all patients. Other studies that consider AS with unique patient classes model independently and distinctly distributed (i.d.d.) service times. The general assumption of independence between the arrival and the service patterns may be questionable. In practice, doctors may increase their service rate, if only subconsciously, during peak hours knowing that there are many patients waiting. This is observed to be the case in a number of studies [4, 52]; ; Rising et al. 1973; Babes and Sarma 1991).

In this thesis, the service time for each patient is considered to be fixed and independent. There is a fixed length of each examination allocated to each patient. Even though the service time may varies from patient to patient, it is assumed that all the allocated services within one day are ought to be finished. Therefore, the modelling of service time is not the main point here.

Queue Discipline is also regarded as the scheduling method. Almost in all studies, it is assumed that arriving patients are served on a first-come, first-served (FCFS) basis. Given punctual patients, this queue discipline is identical to serving patients in the order of their appointment times. However, unpunctuality may cause changes in the actual order of seeing patients, as doctors would not keep idle waiting for the next appointment in the presence of other waiting patients.

In OSC, patients have their urgent levels based on their arriving days. Therefore, a priority rule is used in the scheduling method to determine the most suitable time slots for each patient according to their urgency level. The general case, which is defined in [12, 51], states that the first priority is given to emergencies, followed by second consultations, then scheduled patients; the lowest priority is given to walk-ins that are seen on a FCFS basis.

2.3 Performance Measures

There are many possible performance measures used in the literature to evaluate appointment systems. Most of them use a function of the time patients spend waiting for their appointment or of the time doctors remain idle. Five main types of performance measures are enumerated by [9] and detailed below:

Time-based measures cover mainly the waiting time of patients and the idle and overtime of doctors and resources. Patient waiting time can be further classified as indirect (or virtual), defined as the time between the request for an appointment and the time scheduled for the appointment, or direct (or captive), defined as the time between the scheduled time of the appointment and the time the consultation actually starts. In situations where direct waiting time is calculated and the arrival time of patients is also considered, a common approach is to use the "true" waiting time, defined as the difference from the time the consultation starts to the latest time between the arrival of the patient and the scheduled appointment time. Idle time can be defined as the amount of time a resource is available but not used. Overtime can be seen as the extra time after normal closing hours that a resource is kept busy.

Cost-based measures are generally a linear mapping of the time-based measures to monetary cost. However, it should be noted that a schedule where many patients have small waiting times may be better than a schedule where one patient has excessive waiting, even if the total waiting time in both schedules is the same [32]. When considering more than one performance measure, it is usually enough to establish a relationship between the costs of each measure in order to make a decision. For example, if the objective is to minimise the direct and indirect waiting times of patients, it is possible to provide the ratio of the cost of direct waiting over the cost of indirect waiting. Estimating this ratio may be easier for the service provider, instead of finding the actual monetary costs of each.

Congestion measures may also give an idea of how "good" a system is. Examples of congestion measures include the average number of patients in the queue in a given time period.

Fairness measures: are usually considered as the degree of uniformity of performance across patients. Examples of fairness measures include measures of the average direct waiting time for each patient in the order of appointments (average waiting time of the first patient, of the second, etc.) [4]

Other measures: may include the average number of patients treated in a clinic session, resource utilisation, and any other measure which does not fit into the classifications above.

2.4 Methodologies in Literature

Cayirli classifies research methodologies in appointment scheduling (AS) literature according to the health-care environment on which they focus, and the assumptions they make in [9]. I mainly introduce three type of methodologies here.

2.4.1 Analysis methodologies

The analytical approaches to the study of AS include queuing theory and mathematical programming methods. The main advantage of this exact methods is that they can guarantee the optimality of the solution found.

Integer and mixed integer programming [46] models are commonly used to approach this or other similar problems. [11] define mathematical models for the scheduling of radiotherapy treatment. The objective in their proposed model is to schedule as many patients as possible in a short period of time (e.g. one week). They consider a block system, where a workday is split into a fixed number of time blocks/slots. The limitation of these mathematical models is that they do not consider all constraints present in real-world radiotherapy scheduling, such as machine eligibility, release dates different from the booking requests and patients who require multiple sessions per day. [9]

As described in [9], authors of [28] present a patient scheduling problem in a medical clinic. Patients have stochastic service times and call in advance to arrange an appointment, which can be scheduled to a specific time slot, such that more than one patient can be assigned to the same slot. The goal is to design an algorithm which decides the time of an appointment at the time the patient calls in order to minimise three objective functions: the mean waiting time, idle time and overtime (referred to as "tardiness" in the paper). A local search is proposed and the authors prove that it finds an optimal solution by proving that the objective function is multi-modular.

2.4.2 Simulation

Simulations are also commonly used to model patient scheduling problems. They can be used to better study each specific case, identify bottlenecks, as well as estimate the effect of proposed changes on the scheduling policy, and evaluate different scheduling algorithms [29].

Cayirli investigate the effect of different sequencing rules in scheduling an ambulatory care service in [10]. The investigated rules include sequencing rules, which define the order in which patients are scheduled in appointment slots, and appointment rules, which define the number of patients assigned to each slot and their length. To evaluate each set of rules, a simulation model is used with real-world data from a healthcare clinic in a New York metropolitan hospital. The authors use patient and doctor-based measures to evaluate each combination of rules and shows that sequencing rules have a much higher impact on performance than appointment rules.

Lev and Caltagirone categorise the problem of patient scheduling in a diagnostic radiology department as a classic job shop machine scheduling problem, and develop a discrete event simulation model of patient flow in [38]. The model is used to evaluate the performance of seven different scheduling rules according to 4 performance measures: waiting time prior to examination, total time in the system, distributions of waiting and total times, and the number of patients in the system at the end of working hours. The two rules prioritise the patients in the queue based on their expected session duration for a resource achieve the best results. The authors recommend one of these two rules. However, they acknowledge that, of the evaluated rules, the two best are the only rules which would require a computer to perform the scheduling (the other rules could be performed manually).

2.4.3 Heuristics and Meta-heuristics

In the case where the problem instances are too large for exact methods, heuristics and meta- heuristics can be used instead. As with exact methods, they should also be combined with different approaches to consider future patients, such as resource reservation or demand forecasting.

Vermeulen presents an adaptive algorithm for scheduling patients on a CT-scan in [58]. Patients are divided in groups according to their urgency and other characteristics, where urgent patients have a much shorter time window to get their scan than other patients. The algorithm makes reservations for each type of patient and adaptively modifies the reserved slots when they are not used by non-urgent patients. This is a good example of resource reservation which is updated as time goes by depending on the quantity of the resource available in the short term.

kapamara uses a steepest hill climbing method for a radiotherapy patient scheduling problem in The Arden Cancer Centre radiotherapy department in Coventry, UK [30]. The schedule is first generated by constructive heuristics, where a different dispatching rule is used for each stage of pre-treatment. This schedule is then improved by a hill climbing heuristic which tries to bring each appointment forward to the earliest day possible. The algorithms consider only patients who have already arrived, which often results in a very full schedule with little or no room for higher priority patients who arrive with short notice.

2.4.4 Demand Estimation and Modelling

Methods that estimate the demand and properly model the problem are also used in many research. For example, Nathan presents a model based on a Monte-Carlo distribution to calculate the percentage of spare capacity required to keep waiting times to treatment short [45]. Some environmental factors such as no treatment on bank holidays are analysed with respect to the outcome of the model. Alexopoulos et. al. propose a modelling strategy for patients arrival in community clinics [2]. They suggest that the usual modelling methods are not very precise such as using a Poisson process for modelling arrival of unscheduled patients and a normal distribution for calculating the tardiness of scheduled patients. They perform experiments with several models and distributions, and conclude the Johnson CDF has a better fit than the normal distribution for the problem.

Chapter 3

The Multiple Examinations Scheduling Problem

In this chapter the Multiple Examinations Scheduling Problem (MESP) is formally described with suitable notation in the first section and a multi-agent system as well as a novel solution (scheduling method) for the MESP, called "SlotRkr", is present in section 3.2. Furthermore, the detailed designs of a simulation model that models the random patient arrival streams in the specialty clinic is demonstrated in section 3.3. Finally, the design of the experiments, which aim at testing the performance of such solution, is described in the last section together with the analysis of the experimental result.

3.1 Formal problem description

The multiple examinations scheduling problem (MESP) consists in the design of an effective plan that contains the scheduled date of two examinations, MRI scan and TRUS BIOPSY test, in real-time. Creating effective plan for patients means the average access time to the examination for all patients can be minimised. The plan is created upon patient arrival, making this problem is an on-line scheduling problem which reflects the control mechanisms that deal with monitoring the process and reacting to unplanned events [25]. To properly describe the MESP, three main sets and their mathematical notation is first introduced as follows:

One-Stop-Clinic (OSC) provides consultations for patients once a week and after a total

$|W|$ weeks' period, which is also called "scheduling horizon", the performance of SlotRkr and other scheduling method is calculated and compared to each other. The patient set P contains all patients visiting OSC during the scheduling horizon. Each patient p_i has an important attribute planning window pw_i, which indicates the number of days left between current day and deadline. Then, patients will be classified into three groups according to their planning window: patients with largest planning window $pw_i \in [15, 21]$ are categorised to g_1; patients with smallest planning window $pw_i \in [0, 7]$ are categorised to g_3 and the rest are categorized to g_2.

Patients from different groups have different urgency level but they are all supposed to undergo two types of examinations, MRI scan and TRUS biopsy test, which are represented as set $E = \{MRI, TRUS\}$ and indexed with e. Each examination can only be performed a certain amount of times on specific working day and this information can be described by a resource calendar RC. Resource calendar for each examination is indexed as rc_e, which describes all the resource for the following $|D|$ days. D is the scheduling horizon, which is discretised in working days d and all the waiting patients should be scheduled to undergo examinations after d days ($d \leq |D|$). Considering the deadline of TRUS biopsy test is 21 days in NHS standard, $|D|$ should be equal to 21. Resource in day d is represented by a list of binary variables $\{ts_1, ts_2, ..., ts_{n_d}\}$ where ts_i indicates the availability of $i - th$ time slots on a given day d and n_d is the total number of time slots on that day. An example of resource calendar for OSC is shown in Table.3.5 and Table.3.6

The expected solution to MESP is can be described as effectively scheduling patient p_i to undergo examination e after sd days. Therefore, the variable sd_{e,p_i} denotes the scheduled date of examination e for patient p_i and it cannot be larger than the scheduling horizon $|D|$. Also, the number of patients who have examination e on a certain day cannot exceed the corresponding capacity described in resource calendar. The effective scheduling here is measured by the overdue rate $OD = \frac{|p_i| \mathbb{1}\{sd_{TRUS,p_i} > pw_i\}}{|P|}$, which is the percentage of patients whose scheduled date of last examination, which is TRUS test, exceed the deadline. In this equation, $|P|$ equals the number of patients who visit OSC during W weeks where W defines as the planning horizon. In the original data set, planning horizon equals 50 weeks but in the simulation model, it is a user-defined parameter that can be adjusted. Other performance metrics include the average days between referral date to diagnosis date, denotes as PL, and the average days between consultation and scheduled date of the last

Sets	Indices	Description
E	e	Examinations that patients need to take ($E = \{MRI, TRUS\}$)
RC	rc_e	Resource calendar of examination e_i, contains the number of slots that allocated to performing e in each working day
D	d	Scheduling horizon is a period that discretised in working days and scheduled date should be within such period ($D \in [0, 21]$)
P	p_i	Patients that arrive within $\|W\|$ weeks for scheduling
$G(pw_i)$	g_i	Groups that patient p_i belongs to, which depends on his planning window pw_i ($G(pw_i) = \{g_1, g_2, g_3\}$)

Table 3.1: Sets Notation

examination, denotes as EL, over the planning horizon.

With previous descriptions, I define MESP as follows:

DEFINITION 1 *MESP The Multiple Examinations Scheduling Problem (MESP) can be defined as the problem of creating a set of effective scheduling plan* $\{i \in \mathbb{N} \mid SP_{p_i} = (sd_{MRI,p_i}, sd_{TRUS,p_i})\}$ *for all patients upon individual arrival that determining the scheduled date of Magnetic Resonance Imaging (MRI) scan and Trans Rectal Ultra Sonography (TRUS) test for each of them so that the ratio of patients who are scheduled outside their deadline for diagnosis can be minimised.*

The hospital environment is highly complex and uncertain. For example, even though the length of each timeslot is set to be fixed and identical to every patient, the actual length of examination may vary due to different physical conditions of patients, resulting in a postpone of future examination. For another example, patients may be late or even not attend the scheduled examination, even though it is less frequent in specialised clinic. Therefore, in order to reduce the uncertainty of the MESP to a suitable level, I present a list of assumptions in Table.3.3.

With these assumptions, the only randomness of MESP remains in the patient set. First, the number of patients arrival in each week is uncertain due to the fact that once patients get referral from GP, they can come to visit OSC on every Tuesday, which may be largely affected by the weather or some unanticipated event or even some seasonal factors

Variables	Domain	Description
λ	$\mathbb{R}$	Mean of a Poisson distribution describing the arrivals of patients in each week
μ_i	$\mathbb{R}$	Mean arrival rate of each group in each week
σ_i^2	$\mathbb{R}$	Variance of the arrival rate of each group in each week
B	$\mathbb{N}$	Number of batches to simulate (generate samples) for experiments
W	$\mathbb{N}$	Number of simulated weeks in a batch during which the performance metrics are calculated
pw_i	$\mathbb{N}$	Planning window for patient p_i, indicates the day between current day and his deadline of diagnosis
$ts^e_{d,j}$	$\{0,1\}, \forall j \leq \lvert RC_e(d)\rvert$	The indicator of whether $j-th$ time slot on day d for examination e is occupied (0) and or not (1)
$TotalPat(b,w)$	$\mathbb{R}$	Total number of patients arriving in week w in $b-th$ simulated batch
sd_{e,p_i}	$\mathbb{N} \leq \lvert D\rvert$	Scheduled date of examination e for patient p_i
$diff$	$\mathbb{R}$	Difference of days of scheduling date for two examinations
$COST$	$\mathbb{R}$	The dynamic cost of timeslots calculated by a cost function
C_R	$\mathbb{N}$	A component in cost function that considering the rules set by NHS UK and PAH
C_U	$\mathbb{N}$	A component in cost function that considering the utilisation of examination timeslots
C_P	$\mathbb{N}$	A component in cost function that considering the penalty for scheduling less urgent patient for early examinations
w_R, w_R, w_R	$\mathbb{N}$	Weights of each component C_R, C_R, C_P in the cost function
OD_M	$[0,1]$	The average overdue rate in $\lvert W\rvert$ weeks considering the maximal acceptance pathway length
OD_G	$[0,1]$	The average overdue rate in $\lvert W\rvert$ weeks considering the good practice pathway length
OT	$[0,1]$	The average overtime rate in $\lvert W\rvert$ weeks
L	$\mathbb{N}$	The average length between referral to diagnosis (diagnosis pathway length) in $\lvert W\rvert weeks$
U	$\mathbb{N}$	The average utilisation rate of examination timeslots in $\lvert W\rvert$ weeks

Table 3.2: Variables Notation

Assumptions for MESP
1. The capacity of each examination is fixed and consistent in each week once it has been set
2. For patients who cannot be scheduled with their deadline, they are assumed to receive overtime service, which is not considered in MESP.
3. Each patient arrives is assumed to receive both examination.
4. No-shows and delays for examinations are not considered.
5. All scheduled examinations are assumed to be completed on the scheduled date.

Table 3.3: Assumptions for MESP

such as the holidays. Furthermore, patients' arriving day can directly determine the groups that patients belong to: the later patients visit the hospital, the less time left for them to complete the examinations before they get diagnosis, leading to an uncertain proportion of each group. As the scheduling decision should be made upon patient's arrival, the uncertainty requires an dynamic scheduling rule which has the ability to "predict" the future rather than a static rule. For an instances, there are two patients p_1 and p_2 coming in sequence for consultation and there is only one time slot left for MRI scan in current week. p_1 has a 14-day's planning window, while p_2's planning window is only 2 days. If patients are scheduled in a "first come first serve" way, p_1 will be scheduled in current week's time slot, resulting an overdue schedule for p_2. If the scheduling rule can "predict" that p_1 may not be so urgent and schedule he to next week's time slot, the efficiency of OSC can be improved. However, when considered another scenario where there are still two patients waiting for scheduling while there are more than two time slots left in current week, it is not wise to schedule p_1 to next week's time slot. Therefore, the main challenge of MESP is to predict these uncertainty and schedule patients in a way that the undesirable effect of the uncertainty can be reduced.

3.2 SlotRkr: A solution to MESP

A multi-agent system solution to MESP is proposed in this section. As stated in [59], multi-agent system can better described the distributed nature of the hospital when there are multiple departments involved in the problem of scheduling a combination of appointments to patients . To properly model the problem, three types of agents are constructed: patient

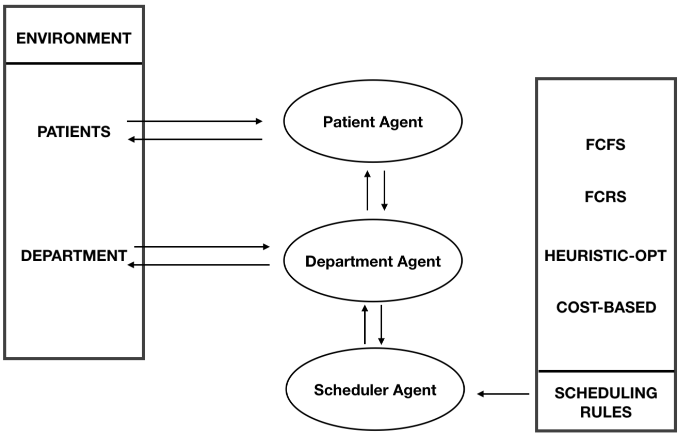

Figure 3.1: Multi-agent System

agent that handles patients' information; department agent that consists of the individual department which provides the examination for patients; and scheduler agent that provides various scheduling rules for scheduling. A novel cost-based scheduling method, called "SlotRkr", is designed and used by scheduler agent. The whole system is shown in Fig.3.1 and the detailed design of each agent as well as their interactions are described below.

3.2.1 Patient Agent

Patient agent is designed to get information needed for scheduling from the electronic patient record (EPR) system, which is shown as the top table in Fig.3.2. The EPR consists of several columns in which the ones interested in MESP include *Pat id, Referral Request Received Date, Date First Seen, Diagnosis Test 1, Ref Date 1, Diagnosis Test 1 - Test, Diagnosis Test 2, Ref Date 2, Diagnosis Test 2 - Test.* Whenever a patient visit OSC, the column *Date First Seen* will be filled by current date. Then, patient agent will extract this line and use it to fill in a new information table that demonstrated in the bottom table in Fig.3.2. Finally, the information contained in the new table will be sent to department agent.

1. A patient counter is used to count the number of patients that come today and the new *Pat id* will be an integer starts from 1;

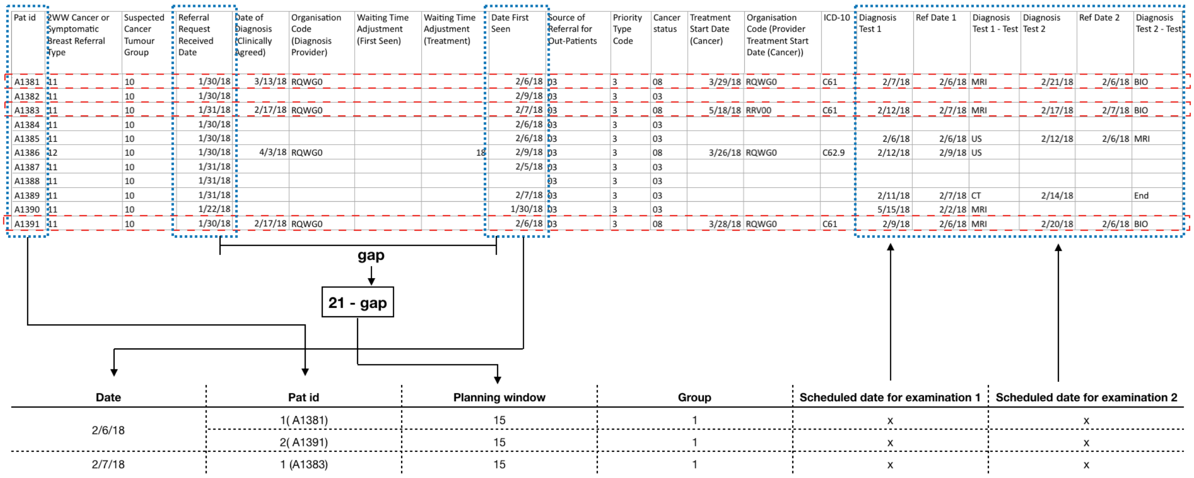

Figure 3.2: Patient agent

2. Planning window is calculated by two steps: first, the gap between referral date and today is calculated, denoted as *gap*; second, the deadline of test finishing date is 21 days after referral so the planning window is $21 - gap$;

3. Patients are assigned to different groups according to planning window;

4. Scheduled date for examinations 1 and 2 are leaving blanked for department agent to fill in.

3.2.2 Department Agent

The capacity plan of OSC determines the total number of timeslots that each department provides on each working day, which can be seen in Table.3.4. Department agent is responsible for maintaining a resource calendar for each department. As an example of resource calendar of MRI that described in Table.3.5, each line represents all the timeslots d days after current day where d is defined in the first column *Days*. The resource calendar will be updated every Tuesday when the consultation is open. Thus, the line with $Days = 0$ represents current day, which is Tuesday and the line with $Days = 7$ represents next Tuesday and so on. On each day, the number of timeslots equals the capacity of that examination on the corresponding day in Table.3.4, which is denoted as n_d. The $i - th$ time slot for MRI scan in day 0 is denoted as $ts^m_{d,i}$ where $i \in [1, n_d]$.

Whenever the time slot is scheduled to one patient, the corresponding position will be updated by the patient's id *Patid* and will be no longer available. When a request of scheduling for examination is received, all first available time slot that on each day are collected as a list and sent to scheduler agent.

Number of time slots		
Weekday	**MRI Department**	**TRUS Department**
Monday	0	5
Tuesday	6	5
Wednesday	2	0
Thursday	0	5
Friday	0	0
Total	8	15

Table 3.4: One Stop Clinic Capacity

3.2.3 Scheduler Agent

Scheduler agent receives two lists of available time slots on different days. First step is to generate all valid combinations of two examinations. A valid combination is defined as two scheduled date of MRI and TRUS in which the scheduled date of MRI is earlier or on the same day of scheduled date of TRUS. The second step is to use one of the scheduling rules to choose one of the combination *comb* from the total set $COMB$. The main contribution

Resource Calendar - MRI						
Days	**Time slots**					
0	$ts^m_{0,1}$	$ts^m_{0,2}$	$ts^m_{0,3}$	$ts^m_{0,4}$	$ts^m_{0,5}$	$ts^m_{0,6}$
1	$ts^m_{1,1}$			$ts^m_{1,2}$		
7	$ts^m_{7,1}$	$ts^m_{7,2}$	$ts^m_{7,3}$	$ts^m_{7,4}$	$ts^m_{7,5}$	$ts^m_{7,6}$
8	$ts^m_{8,1}$			$ts^m_{8,2}$		
14	$ts^m_{14,1}$	$ts^m_{14,2}$	$ts^m_{14,3}$	$ts^m_{14,4}$	$ts^m_{14,5}$	$ts^m_{14,6}$
15	$ts^m_{15,1}$			$ts^m_{15,2}$		

Table 3.5: Resource Calendar - MRI

Resource Calendar - TRUS					
Days	**Time slots**				
0	$ts^t_{0,1}$	$ts^t_{0,2}$	$ts^t_{0,3}$	$ts^t_{0,4}$	$ts^t_{0,5}$
2	$ts^t_{2,1}$	$ts^t_{2,2}$	$ts^t_{2,3}$	$ts^t_{2,4}$	$ts^t_{2,5}$
6	$ts^t_{6,1}$	$ts^t_{6,2}$	$ts^t_{6,3}$	$ts^t_{6,4}$	$ts^t_{6,5}$
7	$ts^t_{7,1}$	$ts^t_{7,2}$	$ts^t_{7,3}$	$ts^t_{7,4}$	$ts^t_{7,5}$
9	$ts^t_{9,1}$	$ts^t_{9,2}$	$ts^t_{9,3}$	$ts^t_{9,4}$	$ts^t_{9,5}$
13	$ts^t_{13,1}$	$ts^t_{13,2}$	$ts^t_{13,3}$	$ts^t_{13,4}$	$ts^t_{13,5}$
14	$ts^t_{14,1}$	$ts^t_{14,2}$	$ts^t_{14,3}$	$ts^t_{14,4}$	$ts^t_{14,5}$
16	$ts^t_{16,1}$	$ts^t_{16,2}$	$ts^t_{16,3}$	$ts^t_{16,4}$	$ts^t_{16,5}$

Table 3.6: Resource Calendar - TRUS

of this thesis is to design a cost-based scheduling rule, called "SlotRkr", to rank all the time slots using a cost function and yield the one with lowest cost. Additionally, a heuristics scheduling rule is designed to approximate the optimal solution of MESP. Two classical scheduling rules, first come first serves (FCFS) and first come random serve (FCRS), are also presented below.

Method 1. Heuristics method

This novel method can be seen as a "postpone and best schedule" method. Its idea is to pretend that the list of waiting patients and their waiting status are known before scheduling. A requirement of this method is a set of patients that is ordered in non-increasing arrival days. All daily fractions from patient are scheduled in the first available feasible examination slots. Instead of scheduling upon arrival, it sorts the list from the most urgent to least and schedules the sorted list in a FCFS sequence. In this way, the most urgent patients will be scheduled to the earliest slots. If the available time slots in current week is all occupied, the unscheduled patients will be added to next week's waiting list. In the next week, the task will be scheduling the newly arrival patients and these unscheduled patients.

Heuristics Method
1. Initialise all parameter $PARA = \{B, W\}$
2. Generate B batches of simulated patient flow $PF_{SIM} = \{P_1, P_2, ..., P_B\}$ where $P_b = \{p_1, p_2, ..., p_W\}$
3. For $b = 1$ to B, do
4. RC_{MRI},$RC_{TRUS} = GetResourceCalendar(W)$
5. For $w = 1$ to W, do
6. $P^{\star}_w = SortByPlanningWindow(PF_{SIM}[b][w])$
7. $weeklyDemand = len(p_w)$
8. For $i = 1$ to weeklyDemand, do
9. $pateint = p_w[i]$
10. $scheduledPlan = FCFS(patient)$
11. Record (P_b)
12. End i
13. End w
14. Calculate $O(P_b)$
15. End b

Table 3.7: Heuristics Method

Even though the method is not practical, it provides an optimal result for comparison and the intuition behind it can be used to build the cost-based function. Details of this method is shown in Table.3.7. Its idea is to assume that the weekly demand is known to the scheduler in advance and then schedule patients from most urgent ones to least urgent ones. In order to achieve this, the weekly demand is first sort by the variable pw in ascending order. Then, scheduling the sorted list in a first-come-first-served manner. Finally, the week that patients are coming (referral week) and the week that they are scheduled to (scheduled week) are record together with patients' planning window. Let variable $T = \{0, 1\}$ indicates whether the referral week of the patient is equal to the scheduled week or not. From the record, we can obtain:

i. the probability of planning window given the $P(pw|T)$;

ii. the probability of transferring $P(T = 1)$ and $P(T = 0)$;

Using Bayesian theorem, the transfer probability for each planning window can be calculated as

$$P(T|pw) = \frac{P(pw|T)P(T)}{P(pw)} = \frac{P(pw|T)P(T)}{P(pw|T = 1)P(T = 1) + P(pw|T = 0)P(T = 0)}$$

, which is integrated to the cost function that described in Method.2. Although this method is not realistic since the scheduling must be made upon patient's arrival and the future patients is not predictable, it can provide an optimal scheduling for the simulated patient flow for comparison.

Method 2. SlotRkr

This method is adapted from [59]. It follows the same idea that defining a schedule-cost function to rank each available timeslot and the one with lowest cost is scheduled to patient. If there are multiple timeslots that have the same lowest cost, the earliest one is scheduled. However, the main difference that distinguishes my work is the re-design of the cost function. The overall cost function consists of three components: C_R, C_U, C_P and each represents the cost of timeslots when considering different performance metrics considered in this problem. It is a dynamic method as the cost function is calculated real-time, which mimics a scheduling staff who make decision while considering the real-time situations. The details are described below.

Component 1: C_R considers the standard rules set by NHS UK and PAH. As mentioned in Chapter 1, it is suggested that *i*) patients should be schedule in the same day or next day for MRI scan after the clinical assessment; *ii*) the scheduled date of TRUS test is 14 days after referral according to PAH; *iii*) the maximal acceptance of TRUS scheduled date is 21 days according to NHS UK. Let binary variables r_1, r_2, r_3 equals 1 if the timeslot does not achieve the corresponding rule. Also, let a_0, a_1, a_2 denotes the parameters that determine the weight of each rule in the cost function. The design rules of the cost function C_R that represents in (3.1) are described below :

i. If $r_3 = 1$, which means the scheduled date of TRUS is later than maximal acceptance due date, the cost should be highest, which is 1.

ii. When $r_3 = 1$, these is no need to consider r_2 as it will always equal to 1. Therefore, when $(1 - r_3) = 1$ and $r_2 = 1$, the scheduled date of TRUS test is between day 14 and 21. In this case, a smaller cost is assigned when compared to the case $r_3 = 1$, which means $a_2 < 1$.

iii. When considering the scheduled date of MRI scan, it is natural to make it closer to the scheduled date of TRUS test. Thus, two scenarios are considered: when $r_2 = 1$ and $r_1 = 0$, larger penalty is added to the cost function, while when $r_2 = 1$ and $r_1 = 1$, a smaller cost is introduced, which means $a_1 > a_0$. The maximal cost of these two scenarios, which is $a_2 + a_1$ cannot exceed 1. Therefore, a intuitive choice used in this thesis is $a_0 = 0.075, a_1 = 0.225, a_2 = 0.375$.

$$C_R = r_3 + (1 - r_3) * \{r_2 * [a_2 + a_1 * (1 - r_1)] + a_0 * r_1\} \tag{3.1}$$

where

$$\begin{cases} a_1 + a_2 < 1 \\ a_0 < a_1 \\ a_0, a_1, a_2 > 0 \end{cases}$$

Component 2: C_U considers the utilisation of timeslots. According to the capacity plan of OSC, there are 15 timeslots for TRUS test per week while there are only 8 timeslots for MRI scan. Furthermore, patients can start TRUS tests only when they finish MRI scan. Due to these reasons, only utilisation of timeslots for MRI scan is considered here. Two variables $FULL$ and $LATE$ are adapted from [59] to calculate the cost function regarding to utilisation. As shown in 3.2 $FULL$ determines the fullness of timeslots in current week; $LATEp_i$ denotes the lateness of the ith patient arriving in current day, which is quite different from [59] where $LATE$ is used to define the lateness of a timeslot in the current day while in our cases, scheduling patients to different timeslots within one day will not affect the objective function. Furthermore, the biggest difference is that in this thesis, utilisation cost function C_U in 3.3 is used to represent a mismatch between the fullness of timeslots and the lateness of patients. A in 3.3, C_U of a timeslot decreases as the lateness of patients increasing with different speed, which is depended on the fullness of timeslots. It has a intuitive meaning: if the timeslots is not full ($FULL \leq \frac{1}{3}$), it is encouraged to schedule these timeslots to patients and the cost sharply reduces for patients with larger lateness as they are more likely to be the last patient arriving in current day; when the fullness of timeslots increased, meaning that there are less available timeslots on a given day, the cost will reduce in a slower speed as patients arrive so as to reserve timeslots for possible higher urgency level patients that coming later.

$$FULL = \frac{\sum_d^6 \sum_i ts^e_{d,i} \mathbb{1}\{ts^e_{d,i} \neq 0\}}{\sum_d^6 \sum_i ts^e_{d,i}}, \qquad LATE_{p_i} = \frac{i}{\lambda} \tag{3.2}$$

$$C_U(LATE_{p_i}, FULL_e) = \begin{cases} (LATEp_i - 1)^2 & \text{if } FULL_e \leq \frac{1}{3} \ , \\ -LATEp_i + 1 & \text{if } \frac{1}{3} < FULL_e \leq \frac{2}{3} \ , \\ -LATEp_i{}^2 + 1 & \text{if } \frac{2}{3} < FULL_e \leq 1 \ , \end{cases} \tag{3.3}$$

Component 3: C_P considers the average rate of overdue patient. Its idea is to penalise the schedule of less urgent patients to early timeslots as such schedule might result in a situation that urgent patients are scheduled to late timeslots which exceed their due date. As discussed in Method 1, a probability of transferring patient to next week's timeslot given patient's planning window $P(T|pw)$ can be obtained by observing the results of Method 1. It is intuitive to treat this probability as a penalty for scheduling patients with pw planning window in current week as the larger probability means it is more likely to postpone such patient into next week's examinations. Therefore,

$$C_P(pw) = P(T = 1|pw) \tag{3.4}$$

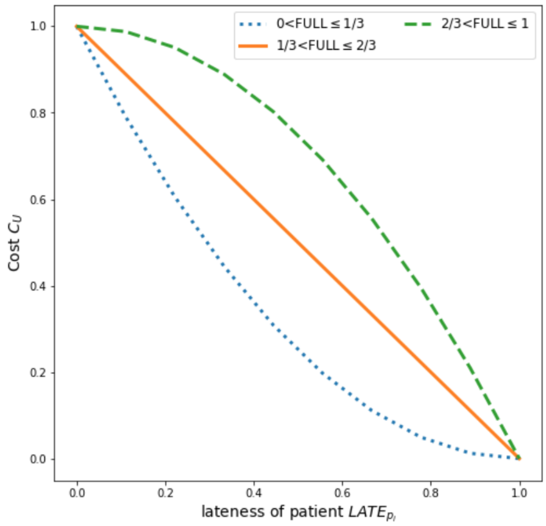

Figure 3.3: Cost C_U

Finally, the overall cost function used in SlotRkr can be represented as

$$COST = w_R C_R + w_R C_U + w_P C_P \tag{3.5}$$

where w_R, w_U, w_P represents the weight of each component and is further tuned in section 4.2.

Method 3. FCFS

When scheduling patients with FCFS, the earliest available timeslot of each department is selected. Given optimal allocation of capacity to patient groups, FCFS is the most efficient static scheduling approach [24]. It serves as the benchmark method in the comparison of different scheduling rules.

3.3 Simulation Model

In this section, a simulation model is developed to model the randomness in the patient's arrival process. It consists of two parts: i) simulate the patient's weekly arrival streams in the OSC; ii) model the composition of the simulated weekly streams.

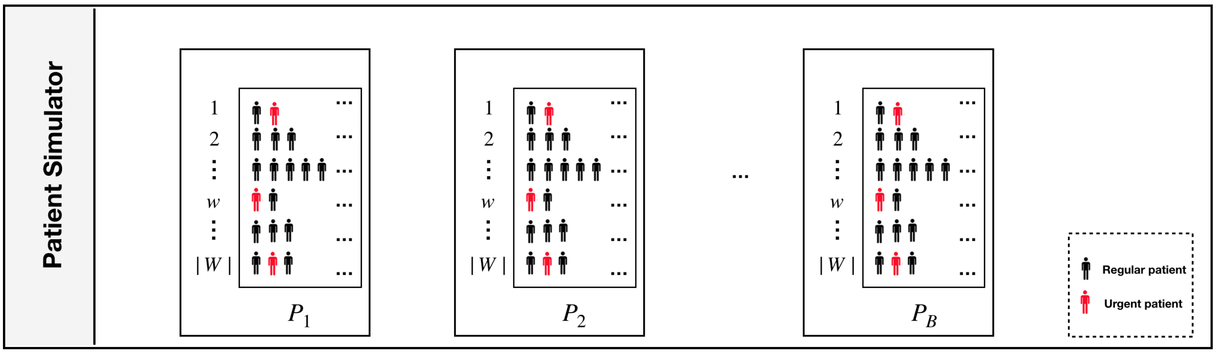

Figure 3.4: Patient simulator

3.3.1 Weekly arrival streams

1. Data Analysis

The first step to model patient's arrival is to examine the data set provided by PAH. The data set consists of in total 2,229 records that describes the patients' activities in Urology clinic from 2017.1 to 2018.8. Among them, there are 837 patients referred to OSC, which is modelled in this thesis, while the rest are referred to other specialty clinic.

The aim is to model the number of patients arriving in OSC for consultation. Let random variable X denotes the number of patients arriving each week and x is the observed value of X. Due to the fact that in OSC, there are only 12 timeslots reserved for consultation each week only on Tuesday and walk-in patients are not acceptable, the level k, or the categories of X, is the integer number ranging from 0 to 12. By grouping patients according to their arriving week, we can obtain the number of arriving patients in each week. Further combing those weeks with same level, we obtain the frequency of weeks of different level, as the bar demonstrate in Fig.3.4.

2. Modelling

Random (unscheduled) patient arrivals were often assumed to follow an ordinary Poisson process (so the corresponding patient interarrival times were randomly sampled from an exponential distribution) [2]. The parameter λ of Poisson distribution represents the average number of arriving patients per week and it can be estimated by the maximum likelihood estimation (MLE):

$$\hat{\lambda}_{MLE} = \frac{\sum_i^n x_i}{n}$$

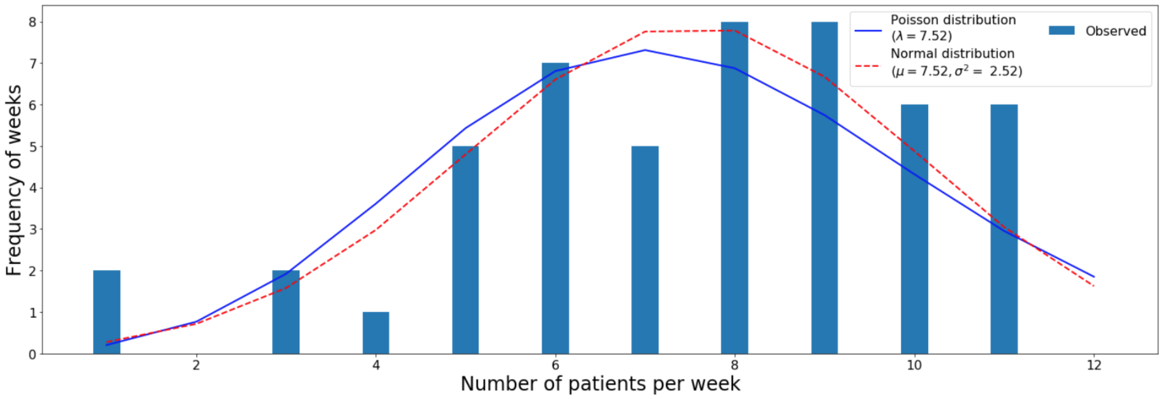

Figure 3.5: Result of simulation model

where n is the total number of weeks in data set and x_i is the observed patient arrivals in ith week. With $\sum_i^n x_i = 376$ and $n = 50$, we have $\hat{\lambda}_{MLE} = 7.52$. For comparison, the normal distribution with parameter $\mu = \sigma^2 = \hat{\lambda}_{MLE}$ is also used to simulate the patients arrivals. The results of poisson simulation and normal simulation can be seen in Fig.3.4 with solid and dash line, respectively.

3. Goodness-Of-Fit Test

In order to examine the accuracy of simulated patient streams from each distribution, two chi-squared goodness-of-fit tests are conducted with the null hypothesis stating that the patient arrivals in OSC are consistent with a Poisson and a Normal distribution, respectively. The main idea of a chi-square goodness-of-fit test is to measure the difference between observed sample frequencies (O_i) and expected frequencies (E_i) by the chi-square statistic $\hat{\chi_c^2} = \sum \frac{(O_i - E_i)^2}{E_i}$. Then, the null hypothesis is rejected if $\hat{\chi_c^2}$ is larger than the value $\chi_\alpha^2(r)$ where r is degrees of freedom, equalling to the number of levels minus 1, and α is the significance level.

Respectively, the calculated $\hat{\chi_c^2}$ of Poisson distribution and Normal distribution is 14.22 and 9.86, which is shown in Table.3.8. With a 95% confidence ($\chi_{0.01}^2(11) = 19.68$), I do not have enough evidence to reject that the arrival of patients follows a Poisson distribution with $\lambda = 7.52$ or a Normal distribution with $\mu = 7.52$, $\sigma^2 = 2.5$. Due to the reason that Poisson distribution is more interpretable when modelling discrete events, it is finally chosen for the simulation of weekly patient arrival streams.

Number of patients each week X	Observed frequencies O_i	Normal distribution frequencies E_i	Poisson distribution frequencies E_i
1	2	0.28	0.20
2	0	0.71	0.77
3	2	1.58	1.92
4	1	2.98	3.61
5	5	4.80	5.43
6	7	6.61	6.81
7	5	7.76	7.31
8	8	7.79	6.88
9	8	6.67	5.74
10	6	4.88	4.32
11	6	3.04	2.95
12	0	1.62	1.85

Table 3.8: Observed and Simulated data

3.3.2 Composition of the streams

As discussed in section 3.1, patients are classified into three groups with different urgency levels according to their planning window. We know OSC provides consultation timeslots every Tuesday for patients with GP referral. However, patients may visit their GP on any workday or even weekend, resulting in different arriving days and the planning window is defined as 21 minus their arriving day, which may introduce lots of randomness. For example, two patients p_1 and p_2 both get referral by their GP on Monday and one patient p_3 get his referral on Wednesday. p_1 visit OSC on the same week, which is one day after he got referral, resulting a 20-days' planning window. p_2 miss the next day's appointment so he and p_3 both visit OSC on next week so their planning window is 13 and 15, respectively. In this case, the composition of the stream can be treated as the percentage of patients with different planning window. Modelling the composition consists of two steps: modelling the composition with respect to group first and modelling the planning window in each group.

1. Group modelling

In the group modelling step, the interested variables are the probability of patients belonging to each group, denoted as $\theta_{G_1}, \theta_{G_2}, \theta_{G_3}$. The MLE of each variable $\hat{\theta}_{G_j}$ can be calculated as dividing the total number of observed patients by the number of observed patients belonging to corresponding group in the data set, which is shown in Equation.3.6. The result of such estimate is $\hat{\theta}_{G_1} = 0.66, \hat{\theta}_{G_2} = 0.29, \hat{\theta}_{G_3} = 0.05$.

$$\hat{\theta}_{G_j} = \frac{\sum p_i \mathbb{1}\{p_i \in G_j\}}{\sum p_i} \tag{3.6}$$

Furthermore, in order to quantify the accuracy of $\hat{\theta}_{G_j}$, the standard error of each estimate should be measured. Since the sample size in the original data set is not enough to verify the accuracy of estimation and it is not possible to generate new samples from it, the bootstrapping technique is used instead. The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method [27]. With the aim of estimating the variability of $\hat{\theta}_{G_j}$, a bootstrapping process is conducted with each step described below:

1. Let F_n denotes the empirical cumulative distribution function of original data set with n samples, and $p^{(i)}\star$ denotes the ith sampling with replacement from the original data set.
2. Set B to a large value, which is 10000 in this bootstrapping.
3. Draw $p^{(1)}\star, ..., p^{(n)}\star \sim \hat{F}_n$
4. Compute bootstrap estimate for θ as $\bar{\theta}^{\star}_{G_1}, \bar{\theta}^{\star}_{G_2}, \bar{\theta}^{\star}_{G_3}$ using the equation in Equation.3.6
5. Repeat step 3 and 4, B times, to get $\{\bar{\theta}^{\star}_{G_1,1}, \bar{\theta}^{\star}_{G_2,1}, \bar{\theta}^{\star}_{G_3,1}, ..., \bar{\theta}^{\star}_{G_1,B}, \bar{\theta}^{\star}_{G_2,B}, \bar{\theta}^{\star}_{G_3,B}\}$
6. The standard error of these bootstrap estimates for group j can be calculated as

$$s.e.boot(G_j) \equiv \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\bar{\theta}^{\star}_{G_j,b} - \frac{1}{B}\sum_{r=1}^{B}\bar{\theta}^{\star}_{G_j,r})^2}$$

The obtained result of bootstrapping standard error for each group is $s.e.boot(G_1) = 0.025, s.e.boot(G_2) = 0.024, s.e.boot(G_3) = 0.011$. With the mean and standard derivation, three beta distributions are constructed to model the probability that patient belongs to different groups.

Finally, according to Theorem 2, it is possible to prove that $\alpha = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu$ and $\beta = \frac{\mu(1-\mu)^2}{\sigma^2} - (1-\mu)$. With $\mu = \hat{\theta}_{MLE}$ and $\sigma^2 = (s.e.boot)^2$, we obtain $\Theta_{G_1} \sim Beta(233.40, 120.17)$,

$\Theta_{G_2} \sim Beta(102.57, 249.38)$ and $\Theta_{G_3} \sim Beta(17.80, 349.80)$. Therefore, whenever a patient arrives, three probability $\theta_{G_1}, \theta_{G_2}, \theta_{G_3}$ are randomly sampled from the corresponding beta distribution and further normalised so that $\theta^\star_{G_1} + \theta^\star_{G_2} + \theta^\star_{G_3} = 1$. Then, this patient's group is defined by a random number r generated from uniform distribution $U[0, 1]$, which is shown in Equation.3.7.

$$Group = \begin{cases} 1 & r \leq \theta^\star_{G_1}, \\ 2 & \theta^\star_{G_1} < r \leq \theta^\star_{G_2}, \\ 3 & \theta^\star_{G_2} < r \leq 1. \end{cases} \tag{3.7}$$

THEOREM 2 *For* $Beta(\alpha, \beta)$ *distribution, we have* $\mu = \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$ *and* $\sigma^2 = Var(X) = \mathbb{E}(X - \mu)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

2. Arriving day modelling

Patients' arrival day refers to the day between referral and first consultation in OSC. Earliest patients that arrive within 7 days after they have referral are grouped into group 1. Group 2 consists of patients with arriving day between 8 and 14 and patients with at least 15-days' arriving are categorised as group 3. The observed frequency of patients in each group is presented in Fig.3.6. According to the data, it can be concluded that:

i. the number of patients in group 1 is the largest (249) and it linearly increases as the arriving day increases;

ii. size of group 2 is less than half of size of group 1 and the number of each arriving day shows no significant difference;

iii. there a peak on the arriving day 18 but the data size is extremely small (18 patients) when comparing to other groups;

iv. there is no patient arriving on day 2,3,15,16 and there are few patients arriving on day 9 and 10, which means there are fewer patients who visit their GP on Saturday or Sunday.

Therefore, three linear regression model is used to predict patients' arrival day in each group and the predicted result is shown as the lines in Fig.3.6. From that linear model, we can obtain the predicted probability of patient' arriving days within each group. Let AR denotes random variables of patients' arriving day and ar is the value that AR can take.

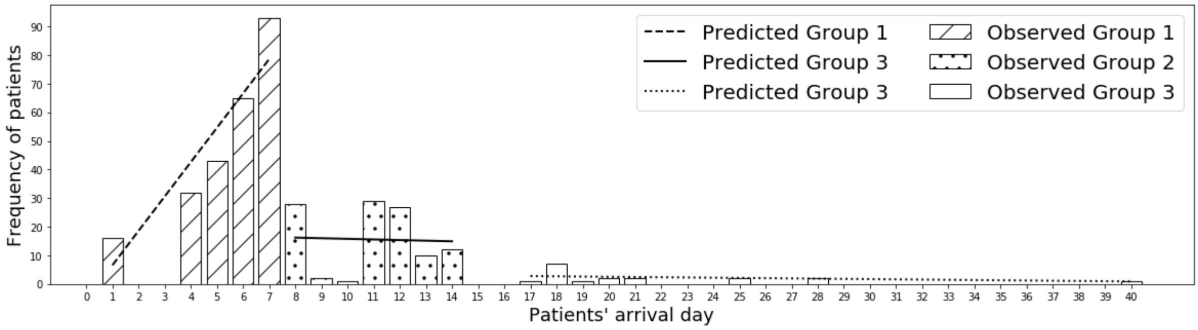

Figure 3.6: Observed and predicted arriving day for each group

The scheduling horizon is 21 days so the maximal value of ar is 21. Furthermore, due to the finding that patients seldom visit GP on weekend, ar equals the integers from 1 to 21, except 2,3,9,10,15 and 16. Multiplying that probability by the sampled probability of the corresponding group θ_{G_i}, we obtain the probability of each arriving days $P(AR = ar)$, which is equal to the probability of patients' planning window $P(pw)$ with planning window $pw = 21 - ar$. Again, by first normalising the probability such that the probability of all valid pw sums up to 1 and then observing the value of a randomly generated number from $U[0, 1]$, we can determine the planning window of each patient in the arrival streams.

3.3.3 Summary

To summarise, the simulation model takes parameter W as input to create W weeks of patient arriving streams. The number of patients per week is generated randomly from a Poisson distribution with parameter $\lambda = 7.52$. Each of them will have a simulated planning window, which is determined by categorising them into three groups first and allocating the planning window within this group. The probability of each group follows beta distribution: $\Theta_{G_1} \sim Beta(233.40, 120.17)$, $\Theta_{G_2} \sim Beta(102.57, 249.38)$ and $\Theta_{G_3} \sim Beta(17.80, 349.80)$.

3.4 Experiments

In this section, a series of experiments are carried out to compare the performance of different scheduling methods under different environmental parameter used in simulation model. First of all, the experimental setting is first introduced. Then, the results of experiments are presented and several hypothesis are tested. Finally, the challenges and limitations of such experiments are discussed.

3.4.1 Experimental setting

1. Test Data

The data used in experiments comes from the simulation model described in section 3.4. During each experiment, there are in total B batches of simulated patient flow which are further divided into W weeks. The patient arrival rate in each week is randomly generated from a Poisson distribution with parameter λ and the probability of each patient belonging to group i is randomly generated from a beta distribution with mean θ_{G_i} and variance $\sigma^2_{G_i}$.

2. Dependent variables

The outcome of the experiments are measured by depend variables described in Table.3.9. Overdue rate is the most importance measurement indicating the performance of scheduling methods. A patient is said to be overdue when his scheduled date for the last examination is exceed the due date and the due date can be determined by two standards: OD_M is the average overdue rate with maximal accepted due date, which is 21 days set by NHS UK, and it can be calculated as

$$OD_M = \frac{\sum_i^N p_i \mathbb{1}\{SD(p_i) > pw_{21}\}}{\sum_i^N p_i}$$

where N is the total number of patients during the scheduling horizon (W weeks), SD_{p_i} is the scheduled date for patient i and pw_M is the planning window calculated according to maximal acceptance. Similarly, OD_G can be calculated in the same way by replacing pw_M with pw_G, which is the planning window calculated according to good practice. Additionally, the variable L can also measure the scheduling efficiency from the perspective of average length of pathway: the longer the length is, the higher probability that patient is overdue. Even though it is similar to the overdue ratio, it is not the main goal considered by PAH.

Apart from overdue rate, overtime rate is another indicator of inefficient scheduling. When there is no available timeslot in the next 21 days, doctors is assumed to work overtime to provide timeslots for patients arriving during this period. The ratio of overtime timeslots to the total timeslots with scheduling horizon is denoted as OT.

The last dependent variable considered is the average utilisation of MRI scan timeslots.

Dependent variables	Description	Weight
OD_M	Average overdue rate with maximal acceptance (21 days)	0.3
OD_G	Average overdue rate with Good practice (14 days)	0.2
OT	Average work overtime rate for doctor	0.2
L	Average length of pathway	0.15
U	Average utilisation of MRI scan timeslots	0.15

Table 3.9: Dependent variables

Although it is not directly considered as the objective, it can still have great impact on scheduling. The reason is that a waste of timeslots could be resulted from a mismatch between the scheduled date and patients' planning window. For example, when the early timeslots on a given week are occupied by less urgent patients who could have used the timeslots in the following days or weeks, the timeslots in the following days or weeks might not be used up when all the following patients arriving have a tight planning window. Therefore, a higher utilisation level means there are fewer timeslots being wasted and it could reflect the efficiency of a scheduling method to a certain degree.

3. Independent variables and Experimental Factors

The independent variables, which are the variables that manipulates in the experiments, include the configurations for simulation model as described in Table.3.10. These variables represent the environmental factors in the hospital. For example, the larger λ will result in a more crowd environment with increased average number of patients per week; and the higher θ_{G_3} will increase the percentage of patients who belong to group 3, which is the most urgent one, resulting in a situation that most patients are requesting the earliest timeslots for examinations. From the data analysis in section 3.3, we obtain the estimated value of each of these parameters as well as the standard error of such estimate, which is presented as a 95%confidence intervals in Table.3.10.

The aim of the experiments is to compare the performance of each scheduling method under different hospital environment and by controlling the values of these independent variables, the environment can be controlled to a certain level. Therefore, two experimental factors are defined as follows.

Factor 1: λ is the patient arrival rate that determines different levels of crowdedness in OSC. Due to the fact that when the clinic is less crowded, it is easier to schedule, I only increase the level of crowdedness from "normal", "crowded" to "extremely crowded", which are identified by the case N, C and EC in Table.3.12, respectively. The value of λ for each level is chosen according to its confidence interval.

Factor 2: $\vec{\theta}$ is a vector containing all the probability of group that patient may belong to. First of all, the case "R" represents a regular level of urgency where the probability of each group is the mean of simulated distribution. By increasing or decreasing the probability of one certain group, there are 6 more different levels of urgency of the weekly demand with full description in Table.3.12. For example, the LG_1 case is determined by increasing the probability of group 1 while others remain the same. After normalisation, the percentage of patients belong to group 2 and 3 will be increased, resulting in a most urgent demand among all the cases. In contrast, HG_1 describes the case that the probability of group 1 is increased while others remain the same, leading to a least urgent demand.

Independent variables	**95% Confidence intervals**
λ	$(6.76, 8.28)$
$\hat{\theta_{G_1}}$	$(0.611, 0.709)$
$\hat{\theta_{G_2}}$	$(0.244, 0.339)$
$\hat{\theta_{G_3}}$	$(0.027, 0.070)$

Table 3.10: Independent variables

Extraneous variables	**Values**
B	30
W	300
w_R	0.461
w_U	0.282
w_P	0.257

Table 3.11: Extraneous variables

Finally, combining the cases of each experimental factor, we obtain $3*7 = 21$ treatments, which are the experimental units. For example, combining case N and R, a treatment identified by $N - R$ can describe a hospital environment with a normal level of crowdedness and a regular level of urgency in the demand. With the aim of finding the most suitable scheduling method that can be used in OSC, it is importance to analyse the performance of each method under treatment $N - R$. Furthermore, it is also interested to investigate the robustness of each method by observing the changes of their performance as the average number of patients or the urgency level of demand increase. Therefore, three hypotheses for the experiments are defined in Table.3.13.

Experimental Factor	**Case ID**	**Levels** (Values)	**Description**
1. Patient arrival rate	N	7.52	The estimated rate $\hat{\lambda}_{MLE}$
	C	8.28	The upper bound of its CI: $\hat{\lambda}_{MLE} + c\frac{\sigma}{\sqrt{n}}$
	EC	9.04	Exceed the upper bound by $c\frac{\sigma}{\sqrt{n}}$
2. Group probability	R	$[0.66, 0.29, 0.05]$	The estimated probability $\hat{\theta}_{MLE}$
	LG_1	$[0.65, 0.30, 0.05]$	The lower bound of its CI: $\theta_{G_1} - c\frac{\sigma}{\sqrt{n}}$
	LG_2	$[0.69, 0.26, 0.05]$	The lower bound of its CI: $\theta_{G_2} - c\frac{\sigma}{\sqrt{n}}$
	LG_3	$[0.67, 0.29, 0.04]$	The lower bound of its CI: $\theta_{G_2} - c\frac{\sigma}{\sqrt{n}}$
	HG_1	$[0.68, 0.28, 0.05]$	The upper bound of its CI: $\theta_{G_1} + c\frac{\sigma}{\sqrt{n}}$
	HG_2	$[0.64, 0.32, 0.05]$	The upper bound of its CI: $\theta_{G_2} + c\frac{\sigma}{\sqrt{n}}$
	HG_3	$[0.65, 0.29, 0.06]$	The upper bound of its CI: $\theta_{G_3} + c\frac{\sigma}{\sqrt{n}}$

Table 3.12: Environmental Setting

Hypotheses for the experiments
1. SlotRkr has better performance, which is defined as having lower weighted performance metrics, than First-come-first-served (FCFS) when the testing scenario best resembles the real environment in OSC
2. SlotRkr has better performance than FCFS when the average number of patients per week increases from normal level to extremely crowded level.
3. SlotRkr has better performance than other scheduling methods when the urgency level of patients weekly demand increases from least urgent to most urgent.

Table 3.13: Hypotheses for the experiments

4. Extraneous variables selection

The extraneous variable is defined as all variables, which are not the independent variable, but could affect the results (depend variables) of the experiment [41]. In this experiments, extraneous variables include the weight of SlotRkr's cost function w_R, w_U, w_P, the number of batches B and the number of weeks in each batch W. Careful selection of each extraneous variable is described below.

1. Number of batches

The number of batches B is a user-defined parameter: a larger B can better determine the variation of results but it is time-consuming. In order to ensure a relative reliable result and a reasonable running time, B is set to 30.

2. Number of weeks

As for the scheduling horizon W, it should be noted the resource calendar is initialised

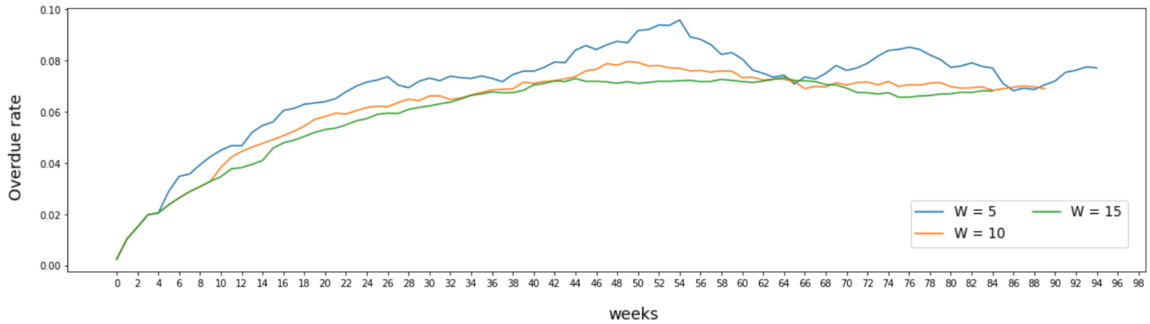

Figure 3.7: Welch's Graphical Method

as an empty one, which may affect the accuracy of experiments as in reality the resource calendar is rarely empty. Therefore, the performance in the early weeks may be much more better than usually and after a period, the performance may reach a steady state. This period is called "warm-up" period. Welch's method is a commonly used method to determine the "warm-up" period [48], which plots moving averages $\bar{X}_j(w)$ of 1 to n observations on a graph for a given time window w [56]. The run length is suggested to be much larger than the warm-up period in order to be able to better capture the steady state performance indicators [35]. The detailed algorithm in [56] is described in Appendix B, and the obtained graph by following such algorithm is shown in Fig.3.7. It is easy to observe that under different setting of time window $w = 5, 10, 15$, the system reaches its steady state around week 30. Therefore, the original scheduling horizon W is extended from 50 to 300 weeks, which is 10 times of the "warm-up" period.

3. Weights

In this subsection, the hyper-parameters in the cost function of SlotRkr are optimised under each case described in Table.3.12 to achieve a better performance. As discussed in Chapter 3, there are three components in the cost function of SlotRkr focusing on different performance metrics. Therefore, the hyper-parameters w_R, w_U, w_P are optimised according to the values of the five performance metrics described in Table.3.9.

Grid search and manual search are the most widely used strategies for hyper-parameter optimization. Grid search is simple to implement and parallelization is trivial; Manual optimization gives researchers some degree of insight into parameter; while empirically and theoretically that randomly chosen trials are more efficient for hyper-parameter optimization than trials on a grid [5]. Therefore, a random search is conducted to search best combination of weights for cost function under different experimental cases defined in 3.12. There are in total 30 runs in each case and in each run, three random values are generated from a $[0, 1]$ uniform distribution to represent w_R, w_U, w_P. These random weights are used in the cost function to scheduling patients generated from the simulation model described in section 3.3 with the selected extraneous variables $B = 2$ and $W = 300$ and the independent variables defined in each case.

The results of random search is analysed using a technique, called "Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)", which is presented by Hwang & Yoon in [26]. Its main idea is to rank each configuration by calculating a "TOPSIS" score which considers all the performance metrics that this configuration achieves. The details of calculating the score can be found in Appendix A. When the score approaches 1, the configuration is said to be maximising the objectives while when it approaches to 0, the configuration tends to better minimising the objectives. Considering the objectives include maximising the utilisation and minimising all the rest metrics, I reverse the utilisation rate to the wasting rate for consistence here. Therefore, the best configuration should be the one with the lowest TOPSIS score.

Finally, the weights with lowest TOPSIS score in different cases are selected and represented together with values of performance metrics in Table.3.14. From top to bottom, the crowdedness levels in the table are sorted from least crowded to most crowded and the urgency levels are sorted from least urgent to most urgent. The bold number is the optimal values obtained among the cases with different crowdedness level; the number in bracket is the optimal values obtained among the cases with different urgency level; the underlined number is the optimal among all cases. By fixing the crowdedness and comparing the optimal weights in different urgency levels, the smallest variance is achieved when urgency level is R. Therefore, the mean of three weights in this urgency levels is finally selected.

Experiment	**Optimal Weights** $\vec{w} = (w_R, w_U, w_P)$	OD_M	OD_G	OT	L	$1-U$	TOPSIS
$N-LG_2$	$\vec{w} = (0.186, 0.068, 0.746)$	0.047	0.448	(0.070)	**13.756**	0.072	0.509
$N-LG_3$	$\vec{w} = (0.535, 0.427, 0.038)$	**0.031**	**0.428**	(0.077)	13.854	(0.013)	0.511
$N-HG_1$	$\vec{w} = (0.431, 0.527, 0.042)$	(0.035)	0.430	(0.080)	13.778	**0.004**	0.510
$N-R$	$\vec{w} = (0.408, 0.263, 0.329)$	0.044	0.444	(0.070)	13.902	0.069	0.507
$N-HG_3$	$\vec{w} = (0.372, 0.154, 0.474)$	0.060	0.463	(0.069)	14.092	0.058	0.508
$N-HG_2$	$\vec{w} = (0.146, 0.588, 0.266)$	0.050	0.468	(0.067)	14.038	0.079	0.510
$N-LG_1$	$\vec{w} = (0.205, 0.194, 0.602)$	0.048	0.466	**(0.062)**	13.963	0.073	0.509
$C-LG_2$	$\vec{w} = (0.881, 0.053, 0.066)$	**(0.036)**	**0.401**	0.114	13.524	0.007	0.509
$C-LG_3$	$\vec{w} = (0.129, 0.536, 0.335)$	0.038	0.442	**0.091**	13.734	0.060	0.507
$C-HG_1$	$\vec{w} = (0.836, 0.085, 0.080)$	0.039	0.412	0.155	**13.330**	**0.003**	0.507
$C-R$	$\vec{w} = (0.450, 0.126, 0.423)$	0.049	0.425	0.137	13.447	0.034	0.505
$C-HG_3$	$\vec{w} = (0.869, 0.112, 0.019)$	0.044	(0.422)	0.148	13.453	(0.004)	0.506
$C-HG_2$	$\vec{w} = (0.651, 0.317, 0.032)$	0.044	0.430	0.136	13.648	0.005	0.506
$C-LG_1$	$\vec{w} = (0.440, 0.533, 0.026)$	0.045	0.438	0.112	13.844	0.008	0.507
$EC-LG_2$	$\vec{w} = (0.806, 0.138, 0.056)$	0.042	**(0.387)**	0.206	**(12.895)**	**(0.000)**	0.505
$EC-LG_3$	$\vec{w} = (0.868, 0.063, 0.069)$	**(0.029)**	(0.410)	0.188	(13.066)	(0.001)	0.506
$EC-HG_1$	$\vec{w} = (0.553, 0.421, 0.027)$	0.040	(0.396)	0.204	(12.958)	**(0.000)**	0.504
$EC-R$	$\vec{w} = (0.525, 0.457, 0.019)$	(0.040)	(0.402)	0.216	(12.906)	**(0.000)**	0.504
$EC-HG_3$	$\vec{w} = (0.099, 0.231, 0.670)$	0.054	0.423	0.198	(13.229)	0.025	0.503
$EC-HG_2$	$\vec{w} = (0.014, 0.472, 0.514)$	(0.040)	(0.417)	0.202	(13.062)	(0.016)	0.503
$EC-LG_1$	$\vec{w} = (0.550, 0.440, 0.010)$	(0.043)	(0.414)	**0.187**	(13.205)	**(0.000)**	0.504

Table 3.14: Results obtained varying the parameter w_R, w_U, w_P in Equation.3.5, where each column represent the value of different performance metric and TOPSIS score

3.4.2 Experimental Results and Discussion

In accordance with the experimental settings described before, there are in total 21 experiments. In each experiment, the patient arrival streams are generated by the simulation model with different independent variables as the input. Then, four scheduling methods are used to create schedule plans for each patient. The depend variables are calculated according to the final schedules that they output. With the obtained results, each hypothesis is examined as follows.

Hypothesis 1

The first hypothesis states that SlotRkr can have a better performance than FCFS when the testing scenario best resembles the real environment in OSC. This testing scenario is denoted as $N-R$ and the performance of the heuristics method, SlotRkr and FCFS are demonstrated in Table.xx.

Hypothesis 2
The second hypothesis states that SlotRkr can have a better performance than FCFS when the average number of patients per week increases from normal level to extremely crowded level. Therefore, I examined the test results of all the urgency levels with different crowdedness level. With a certain urgency level, whenever the crowdedness level increases (e.g. from N to C and from C to EC), the absolutely difference between the scores obtained by each method is calculated and shown in Table.3.15. From that table we can clearly observed that when the crowdedness level increases, only SlotRkr has a negative absolutely difference, indicating that the score is decreased as clinic becomes more crowded. It should be noted that the score is calculated by the weighted values of each objectives, which is optimal when it is minimised. Therefore, I do not reject Hypothesis 2.

Hypothesis 3
The third hypothesis states that SlotRkr can have a better performance than FCFS when the urgency level of patients weekly demand increases from least urgent to most urgent. Similarly, I examined the test results of all the crowdedness levels with different urgency level. Following the same steps described above, I generate a new table from the experiments results as shown in Table.3.16. Read it from top to bottom, we can see that... Therefore, I do not reject Hypothesis 3.

Urgency level	Absolutely Difference of Score					
	Heuristics		**SlotRkr**		**FCFS**	
	N→C	C→EC	N→C	C→EC	N→C	C→EC
xLG_2	5.71%	1.76%	-1.39%	-1.41%	5.10%	1.27%
LG_3	5.62%	1.94%	-0.92%	-1.36%	4.96%	1.39%
HG_1	5.62%	1.94%	-0.92%	-1.36%	4.96%	1.39%
R	5.62%	1.94%	-0.92%	-1.36%	4.96%	1.39%
xHG_3	5.98%	1.61%	-1.00%	-1.20%	5.33%	1.09%
HG_2	5.62%	1.94%	-0.92%	-1.36%	4.96%	1.39%
xLG_1	6.48%	1.46%	-0.75%	-1.60%	5.85%	0.88%

Table 3.15: Results obtained by varying the crowdedness levels with fix urgency level

Urgency changes	Absolutely Difference of Score								
	N			**C**			**EC**		
	H	S	F	H	S	F	H	S	F
$LG_1 \rightarrow HG_2$	1	2	3	1	2	3	1	2	3
$HG_2 \rightarrow HG_3$	1	2	3	1	2	3	1	2	3
$HG_3 \rightarrow R$	1	2	3	1	2	3	1	2	3
$R \rightarrow HG_1$	1	2	3	1	2	3	1	2	3
$HG_1 \rightarrow LG_3$	1	2	3	1	2	3	1	2	3
$LG_3 \rightarrow LG_2$	1	2	3	1	2	3	1	2	3

Table 3.16: Results obtained by varying the urgency levels with fix crowdedness level

Chapter 4

Conclusion and Future Work

In this chapter, the main contributions of all previous chapters are first summarised in the section 4.1. In that section, the problem background as well as the motivations and objectives of this thesis are briefly reviewed, followed by a summary of reviewed literature. Finally, the contributions of proposed solution to the problem are listed. In section 4.2, the future direction of research is proposed.

4.1 Summary of Contributions

This thesis focuses on the multi-stage online scheduling problem, which remains a key challenge in the prostate cancer pathway faced by many hospitals. Patients' random arrival process introduces uncertain factors that could largely affect the efficiency of scheduling. Different resource calendars are maintained by different departments and the coordination between the timeslots provided by them is essential. Furthermore, an online schedule requires the scheduling system to dynamically consider the status of incoming patients and the available resource when making schedule, making this problem even more challenge.

For a better understanding of the background of the problem investigated in this thesis, the available literature on outpatient scheduling is classified according to the environment they deal with, different characteristics of the pattern of patient arrivals, service times and preferences of patient and provider, and the performance measures they considered. Also, methodologies in literatures are categorised into three types: analysis, simulation and demand estimation. The advantages and limitations are discussed along with the corresponding literatures and the solution in this thesis follows a simulation methodology.

In Chapter 3, an detailed description about the problem is first demonstrated. With a fully understanding of the background of the problem, the key challenges are first identified as follows: *i*) each patient has a different and independent start date of pathway when they are referred to PAH by their GP and with the same maximal accepted length of pathway set by NHS UK, each of them will have a different due date of pathway; *ii*) patients' arrival process is random and the number of patients arriving per week is not determined. Then, importance events during scheduling process are classified and mathematical represented as *Sets* and *Variables* in Table.3.1 and Table.3.2, respectively. At the end of section 3.1, a formal definition of Multiple Examinations Scheduling Problem (MESP) is given and three reasonable assumptions about it are made to refine the focus of this problem.

The main contribution in Chapter 3 is to present a comprehensive framework to address the Multiple Examinations Scheduling Problem (MESP). One important part of this framework is the simulation model which captures the randomness in the patient's arrival process. By analysing the historical patient record from 2017 to 2018 provided by PAH, a Poisson distribution and three beta distributions are constructed to model the randomness in patient's arrival process. To be more specified, I first uses Poisson and Normal distribution to simulate the weekly patient arrivals. With a Chi-Square Goodness-Of-Fit test, I concluded that both distributions can simulate the real distribution with a 95% confidence and Poisson distribution is finally used. The other important work in this section is to predict the distribution of patients' arriving day, which is the day between patient's referral date and the first appointments date. By first categorising the arriving days into three groups, three beta distributions are constructed with the MLE mean and the standard error estimated through bootstrapping. Then, the arriving days in each group are modelled by three linear regression models. The advantages of this simulation model is that all these parameter used in the simulated distributions can be further adjusted to simulate different environment, which can be further used to generate experimental data. The main limitation is that original data set that the linear regression model based on is not big enough to ensure a generalised result. Therefore, the simulated patient streams may highly resemble the historical data.

The other key part in the framework is the multi-agent system that designed to model the activities between patients, department and scheduler staff in the clinic. Similar to [59], patient agent and department agent are developed. Patient agent is responsible for extracting schedule-related information from patient's record; department agent maintains a resource calendar of the timeslots for each examinations. Additionally, I create a scheduler agent which can make combinational schedule plan for two examinations. The main contribution in this part is the development of two scheduling methods that can be used by scheduler agent to make plan: the heuristics method and a cost-based method, called "SlotRkr". With the heuristics method, the optimal scheduling can be achieved during the scheduling horizon as it uses future information when making schedule plans. Even though it is not realistic in practice, it provides a transferring probability with respect to each planning window for SlotRkr. With that probability, SlotRkr aims at dynamically calculating the cost for combinational timeslots and yielding the one with lowest cost. As one of the advantages, the design of components in the cost function is intuitive as it considers the current resource calendar and predicts future demands, which mimics human schedulers' consideration when they need to trade-off between the risk of no early timeslot available for urgent patients and the waste of timeslots if they are reserved and not being used. However, the limitation may remain in the inaccurate simulation of real situation as the data size is too small to find a generate solution.

Finally, a series of experiments are carried out to test the performance of each scheduling method under different testing scenarios. Test data, or the simulated patient streams during a certain scheduling horizon, is randomly generated by the simulation model and the parameters of the model are considered to be independent variables, which are carefully controlled to simulate different testing scenarios. Depend variables include the overdue rate considering a maximal accepted pathway length (21 days) and a good practice pathway length (14 days) set by NHS UK and PAH, respectively. Apart from that, other factors, including overtime rate, average pathway length and machine utilisation rate, that may indirectly reflect the goal are also considered. The values of extraneous variables are selected in a way that their impact to the results is minimised. Result of such experiments shows that the hypotheses 1 and 2 is not rejected.

4.2 Future Work

The future research directions inspired by this work remains in two aspects. First, the components in the cost function in SlotRkr can be further improved to *i*) consider more complex objectives, such as patients' preference discussed in [22, 59]; *ii*) model the considered objectives in a more accurate way, such as using the non-linear cost function instead of a linear one. Also, the multi-agent system can be designed in a more efficient way to increase the scheduling speed, which could be an important indicator in an online scheduling system.

The second aspect focuses on adopting more methodologies to schedule. For example, resource reservation is proved to be useful in when patients are classified into different categories [48]. Re-scheduling or conflict resolution after an initial scheduling are common methods in a multi-agent system [14, 59].

Appendix A

TOPSIS

Its first step is to find the ideal and negative-ideal configuration, which achieve the best and worse objective functions among all the configurations. Then, it calculate a TOPSIS score of each configuration by measuring the distance between the ideal and negative-ideal configuration. In order to be more self-maintained, I briefly describe the calculation of such score below.

Let's assume there are in total M objective functions and N configurations, and the average value achieved for objective function m with configuration n in several simulations is denoted as v_{nm}, which are stored in a decision matrix of M columns and N rows. Let a m length vector stores the weight of each objective function, denoted as ϵ_m. Therefore, a normalised decision matrix can be calculated as

$$v_{nm}^{'} = \varepsilon_m \frac{v_{nm}}{\sqrt{\sum_{n'=1}^{N} v_{n'm}^2}} \tag{A.1}$$

where the weight ε_m are user-defined such that $\sum_{m=1}^{M} \varepsilon = 1$. The ideal and negative-ideal configuration for objective function m are the the largest and smallest values in the matrix, denoted as $v_m^{'\star}$ and $v_m^{'-}$, respectively. Therefore, the distances between each values in matrix and the ideal and negative-ideal values can be calculated as

$$D_n^\star = \sqrt{\sum_{m=1}^{M} (v'_{nm} - v'^{\star}_m)^2} \tag{4.4}$$

and

$$D_n^- = \sqrt{\sum_{m=1}^{M} (v'_{nm} - v'^{-}_{m})^2} \tag{A.2}$$

, respectively. Finally, the TOPSIS score of configuration n is given by

$$D_n = \frac{D_n^-}{D_n^\star + D_n^-} \tag{A.3}$$

Appendix B

Statistical Results

Experi -ment	Avg. OD_M			Avg. OD_G			Avg. OT			Avg. L			Avg. U			Score		
	O	**S**	**F**	**O**	**S**	**F**	**O**	**S**	**F**	**O**	**S**	**F**	**O**	**S**	**F**	**O**	**S**	**F**
$N-LG_1$	2.29	3.17	4.23	29.07	39.55	30.42	7.64	10.84	7.64	12	13	12	98.36	94.96	98.36	1.98	2.27	1.99
$N-HG_2$	2.03	2.98	3.99	29.66	39.96	31.29	6.6	9.46	6.6	12	13	12	98.42	95.04	98.42	1.99	2.27	2.0
$N-HG_3$	2.05	3.03	3.95	30.26	40.17	31.24	7.84	10.99	7.84	12	13	12	98.51	95.14	98.51	2.0	2.27	2.0
$N-R$	2.12	3.03	3.98	28.18	38.88	29.59	6.91	9.77	6.91	12	13	12	98.35	94.98	98.35	1.97	2.25	1.98
$N-HG_1$	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$N-LG_3$	1.54	2.38	3.1	26.79	38.04	28.68	7.56	10.87	7.56	11	13	11	98.29	94.75	98.29	1.94	2.24	1.95
$N-LG_2$	2.3	3.17	4.11	26.8	37.69	28.51	7.8	11.11	7.8	12	13	12	98.32	94.78	98.32	1.95	2.24	1.96
$C-LG_1$	3.29	3.56	4.88	39.64	39.87	35.51	13.89	15.76	13.89	12	13	12	99.66	97.53	99.66	2.11	2.25	2.11
$C-HG_2$	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$C-HG_3$	3.84	4.07	5.42	39.13	39.68	35.13	13.94	15.92	13.94	12	13	12	99.65	97.36	99.65	2.11	2.25	2.11
$C-R$	3.13	3.3	4.58	38.8	38.99	34.58	12.93	14.58	12.93	12	13	12	99.64	97.49	99.64	2.09	2.23	2.09
$C-HG_1$	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$C-LG_3$	2.25	2.59	3.75	37.3	38.38	33.74	13.61	15.64	13.61	12	13	12	99.61	97.28	99.61	2.07	2.22	2.07
$C-LG_2$	3.19	3.38	4.59	36.13	37.48	32.71	14.27	16.29	14.27	12	13	12	99.58	97.25	99.58	2.06	2.21	2.06
$EC-LG_1$	3.92	3.66	4.88	46.28	39.33	36.9	20.23	21.16	20.23	12	13	12	99.92	98.76	99.92	2.14	2.21	2.13
$EC-HG_2$	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$EC-HG_3$	4.77	4.31	5.6	45.81	39.15	36.72	20.39	21.3	20.39	12	13	12	99.93	98.8	99.93	2.15	2.22	2.13
$EC-R$	3.73	3.51	4.73	45.09	38.23	35.86	20.66	21.57	20.66	12	13	12	99.94	98.81	99.94	2.12	2.19	2.11
$EC-HG_1$	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
$EC-LG_3$	22.82	2.75	3.91	43.88	37.43	34.84	20.44	21.39	20.44	12	13	12	99.92	98.73	99.92	2.1	2.17	2.08
$EC-LG_2$	3.88	3.58	4.8	42.19	37.29	34.55	20.11	21.09	20.11	12	13	12	99.94	98.71	99.94	2.1	2.18	2.09

Bibliography

[1] *Scheduling algorithms*. Springer, 2007.

[2] Christos Alexopoulos, David Goldsman, John Fontanesi, David Kopald, and James R Wilson. Modeling patient arrivals in community clinics. *Omega*, 2008.

[3] Malika Babes and GV Sarma. Out-patient queues at the ibn-rochd health centre. *Journal of the Operational Research Society*, 1991.

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

[5] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 2012.

[6] Ingeborg A Bikker, Nikky Kortbeek, Rob M van Os, and Richard J Boucherie. Reducing access times for radiation treatment by aligning the doctor's schemes. *Operations research for health care*, 2015.

[7] M Brahimi and DJ Worthington. Queueing models for out-patient appointment systems—a case study. *Journal of the Operational Research Society*, 1991.

[8] Brecht Cardoen, Erik Demeulemeester, and Jeroen Beliën. Operating room planning and scheduling: A literature review. *European journal of operational research*, 2010.

[9] Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: a review of literature. *Production and operations management*, 2003.

[10] Tugba Cayirli, Emre Veral, and Harry Rosen. Designing appointment scheduling systems for ambulatory care services. *Health care management science*, 2006.

[11] Domenico Conforti, Francesca Guerriero, and Rosita Guido. Optimization models for radiotherapy patient scheduling. *4OR*, 2008.

[12] Trevor F Cox, John P Birchall, and Henry Wong. Optimising the queuing system for an ear, nose and throat outpatient clinic. *Journal of Applied Statistics*, 1985.

[13] Robert W Day, Matthew D Dean, Robert Garfinkel, and Steven Thompson. Improving patient flow in a hospital through dynamic allocation of cardiac diagnostic testing time slots. *Decision Support Systems*, 2010.

[14] Keith Decker and Jinjiang Li. Coordinating mutually exclusive resources using gpgp. *Autonomous Agents and Multi-Agent Systems*, 2000.

[15] Richard A Deyo and Thomas S Inui. Dropouts and broken appointments: a literature review and agenda for future research. *Medical Care*, 1980.

[16] NHS England. 2017/19 nhs standard contract: November 2016. https://www.england.nhs.uk/nhs-standard-contract/17-18/, November 2016.

[17] NHS England. Implementing a timed prostate cancer diagnostic pathway. https://www.england.nhs.uk/wp-content/uploads/2018/04/implementing-timed-prostate-cancer-diagnostic-pathway.pdf, April 2018.

[18] NHS England. Nhs 'one stop shop' for prostate cancer means faster and more accurate diagnosis. https://www.england.nhs.uk/2018/03/nhs-one-stop-shop-for-prostate-cancer-means-faster-and-more-accurate-diagnosis/, March 2018.

[19] Robert B Fetter and John D Thompson. Patients' waiting time and doctors' idle time in the outpatient setting. *Health services research*, 1966.

[20] Linda V Green, Sergei Savin, and Ben Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 2006.

[21] Diwakar Gupta and Brian Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 2008.

[22] Diwakar Gupta and Lei Wang. Revenue management for a primary-care clinic in the presence of patient choice. *Operations Research*, 2008.

[23] Chrwan-Jyh Ho and Hon-Shiang Lau. Minimizing total cost in scheduling outpatient appointments. *Management science*, 1992.

[24] Wallace J Hopp and Mark L Spearman. Factory physics: Foundations of manufacturing management, burr ridge, il. Irwin/McGraw-Hill, 2000.

[25] Peter JH Hulshof. *Integrated decision making in healthcare*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2013.

[26] Ching-Lai Hwang and Kwangsun Yoon. Methods for multiple attribute decision making. In *Multiple attribute decision making*. Springer, 1981.

[27] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.

[28] Guido C Kaandorp and Ger Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 2007.

[29] T Kapamara, K Sheibani, D Petrovic, O Haas, and C Reeves. A simulation of a radiotherapy treatment system: A case study of a local cancer centre. In *Proceedings of ORP3 Conference, Guimarães*, 2007.

[30] Truword Kapamara and Dobrila Petrovic. A heuristics and steepest hill climbing method to scheduling radiotherapy patients. In *Proceedings of the 35th International Conference on Operational Research Applied to Health Services, Catholic University of Leuven, Belgium*, 2009.

[31] Seongmoon Kim and Ronald E Giachetti. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and humans*, 2006.

[32] Kenneth J Klassen and Thomas R Rohleder. Scheduling outpatient appointments in a dynamic environment. *Journal of operations Management*, 1996.

[33] Linda R LaGanga and Stephen R Lawrence. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences*, 2007.

[34] Linda R LaGanga and Stephen R Lawrence. Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management*, 2012.

[35] Averill M Law, W David Kelton, and W David Kelton. *Simulation modeling and analysis*, volume 2. McGraw-Hill New York, 1991.

[36] B Lehaney, SA Clarke, and RJ Paul. A case of an intervention in an outpatients department. *Journal of the Operational Research Society*, 1999.

[37] Joseph YT Leung. *Handbook of scheduling: algorithms, models, and performance analysis*. CRC Press, 2004.

[38] Benjamin Lev and Robert J Caltagirone. Evaluation of various scheduling disciplines by computer systems. In *Proceedings of the 7th conference on Winter simulation-Volume 1*, 1974.

[39] James M Magerlein and James B Martin. Surgical demand scheduling: a review. *Health services research*, 1978.

[40] Joren Marynissen and Erik Demeulemeester. Literature review on multi-appointment scheduling problems in hospitals. *European Journal of Operational Research*, 2018.

[41] S. A. McLeod. Experimental method. 2012.

[42] A Mercer. A queueing problem in which the arrival times of the customers are scheduled. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1960.

[43] Wojtek Michalowski, Szymon Wilk, Anthony Thijssen, and Mingmei Li. Using a bayesian belief network model to categorize length of stay for radical prostatectomy patients. *Health care management science*, 2006.

[44] Kumar Muthuraman and Mark Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *Iie Transactions*, 2008.

[45] RP Nathan, PJ Thomas, M Jain, AS Murray, and EJ Rhodes. Environmental dose rate heterogeneity of beta radiation and its implications for luminescence dating: Monte carlo modelling and experimental validation. *Radiation Measurements*, 2003.

[46] George L Nemhauser and Laurence A Wolsey. Chapter vi integer programming. *Handbooks in Operations Research and Management Science*, 1989.

[47] Jonathan Patrick, Martin L Puterman, and Maurice Queyranne. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations research*, 2008.

[48] Arturo E Perez Rivera. Proart: Preventing delays via proactive linac-capacity planning. Master's thesis, University of Twente, 2012.

[49] Best Practice. Prostate cancer uk best practice pathway.

[50] Edward J Rising, Robert Baron, and Barry Averill. A systems analysis of a university-health-service outpatient clinic. *Operations Research*, 1973.

[51] Edward J Rising, Robert Baron, and Barry Averill. A systems analysis of a university-health-service outpatient clinic. *Operations Research*, 1973.

[52] John F Rockart and Paul B Hofmann. Physician and patient behavior under different scheduling systems in a hospital outpatient department. *Medical Care*, 1969.

[53] Gordon Swartzman. The patient arrival process in hospitals: statistical analysis. *Health services research*, 1970.

[54] James R Swisher, Sheldon H Jacobson, J Brian Jun, and Osman Balci. Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & operations research*, 2001.

[55] Van-Anh Truong. Optimal advance scheduling. *Management Science*, 2015.

[56] Dr.Çağatay ÜNDEĞER. Output data analysis, 2008.

[57] Nico M Van Dijk. To pool or not to pool?" the benefits of combining queuing and simulation". *Simulation Conference, 2002. Proceedings of the Winter*, 2002.

[58] Ivan Vermeulen, Sander Bohte, Koye Somefun, and Han La Poutré. Multi-agent pareto appointment exchanging in hospital patient scheduling. *Service Oriented Computing and Applications*, 2007.

[59] Ivan B Vermeulen, Sander M Bohte, Sylvia G Elkhuizen, Piet JM Bakker, and Han La Poutré. Decentralized online scheduling of combination-appointments in hospitals. In *ICAPS*, 2008.

[60] Eduardo L Villegas. Outpatient appointment system saves time for patients and doctor. *Hospitals*, 1967.

[61] Quinn E Whiting-O'Keefe, Donald W Simborg, Wallace V Epstein, and Anne Warger. A computerized summary medical record system can provide more information than the standard medical record. *Jama*, 1985.

[62] Jingyu Zhang, Brian T Denton, Hari Balasubramanian, Nilay D Shah, and Brant A Inman. Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 2012.