The Pennsylvania State University
The Graduate School
Department of Industrial and Manufacturing Engineering

ANALYTICS-DRIVEN DESIGN OF MULTI-PHASE MULTI-PROVIDER APPOINTMENT SYSTEM FOR PATIENT SCHEDULING

A Dissertation in
Industrial Engineering and Operations Research
by
Sharan Srinivas

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2017

The dissertation of Sharan Srinivas was reviewed and approved* by the following:

A. Ravi Ravindran
Professor of Industrial Engineering, Pennsylvania State University
Dissertation Adviser and Chair of Committee

M. Jeya Chandra
Professor of Industrial Engineering, Pennsylvania State University

Vittaldas Prabhu
Professor of Industrial Engineering, Pennsylvania State University

Akhil Kumar
Professor of Information Systems, Pennsylvania State University

Janis P. Terpenny
Professor of Industrial Engineering, Pennsylvania State University
Head of the Department of Industrial and Manufacturing Engineering

*Signatures are on file in the Graduate School.

Abstract

Outpatient scheduling plays a key role in matching the healthcare provider capacity to patient demand and improving clinic performance measures, such as patient waiting time, patient satisfaction, and resource utilization. In addition to the traditional pre-booked appointments, outpatient hospitals and clinics are also experimenting with same day appointments. Designing a hybrid appointment system (combination of same-day and pre-booked) involves multiple decisions such as determining the appointment types, patient sequence, and appointment time. Further, various factors such as patient flow, demand uncertainty, and patient no-shows (patients who do not arrive for scheduled appointments) must be considered to develop an effective design. Inefficiencies in the appointment system design and patient no-shows cost the U.S. healthcare system more than $150 billion a year. In addition, they also reduce productivity and timely access to care.

Most of the previous work on outpatient appointment systems consider a simplified clinic setting with single phase (one-stop service) and single provider. Further, they rarely consider patient's provider preference, patient availability, patient specific no-show rate, uncertainty in patient demand and service times. However, in practice, most outpatient departments have multi-phase settings (e.g., pre-screening, visit nurse, visit doctor, check-out) with multiple providers. A detailed simulation analysis indicated that ignoring the multi-phase nature of patient flow, patient's provider preference and patient's availability

lead to unmet demand, patient dissatisfaction and inefficient resource utilization. Further, the associated uncertainties complicate the task of designing the appointment system. This research focuses on designing a data-driven multi-phase multi-provider appointment system for outpatient clinics with the objective of improving resource utilization and patient satisfaction.

First, a new approach to design a hybrid appointment system, a combination of pre-booking and open access (same day) appointment types, is proposed. The objective is to determine the schedule configuration of a hybrid appointment system under uncertainty for a multi-phase multi-provider clinic that incorporates patient's provider preference and availability. A mathematical programming model is proposed to determine the optimal percentage of appointments reserved for pre-booking and open access, and a scenario-based Monte Carlo approach is used to account for uncertainty. Finally, heuristics are developed to determine the best configuration for the hybrid appointment system.

Next, a new framework for sequentially scheduling patients is proposed by using a combination of data analytics and simulation. In the proposed framework, patient-related data from various sources are used to develop predictive models to identify the risk of patient no-show. Finally, different scheduling rules, that leverage the patient specific no-show risk are proposed. Their effectiveness is evaluated with respect to current scheduling practices. The results indicate that the proposed rules consistently outperform the current practice for all the clinic settings tested.

A case study with real data from a Family Medicine Clinic in Pennsylvania is used to show the feasibility and applicability of the proposed models. The analysis of the results provided several key insights in designing an appointment system, which are applicable to both researchers and practitioners. Further, the proposed approaches are generic and can be adopted by any outpatient clinic by incorporating their clinic parameters, such as operating hours, slot duration and others.

Table of Contents

List of Figures

List of Tables

Acknowledgments

This research would not have been possible without the encouragement, support, love and advice of many people. I take this opportunity to extend my gratitude and appreciation to all those who helped me directly or indirectly to complete this dissertation.

First and foremost, I would like to thank God for giving me the strength, patience, perseverance and right opportunity at the right time throughout my Ph.D. program. Next, I would like to thank and express my deep gratitude to Dr. A. Ravi Ravindran, my adviser, for his support, guidance, encouragement and technical expertise under which I was able to perform in-depth research. He has given me many rewarding professional development opportunities. I learned the skills and qualities of a good human and a leader under his guidance.

I thank my dissertation committee members, Dr. M. Jeya Chandra, Dr. Vittal Prabhu and Dr. Akhil Kumar, for providing continual feedback and for serving on my committee. Special thanks to Dr. Prabhu for giving me the opportunity to teach and mentor students. I am extremely grateful to the Department of Industrial and Manufacturing Engineering at the Pennsylvania State University for awarding me the fellowship for the 2016-17 academic year. The fellowship allowed me to spend more time on research with no financial consequences. I thank Penn State College of Medicine and Hershey Medical Center for providing me the data required to test the application of the models proposed in this research. Special thanks to Mr. Matt Bolton of Hershey Medical Center for his efforts in data collection and secure transmission.

I thank my amazing wife, Suchithra, for her love, support and patience. She has been my greatest supporter and strongest critic. I would like to thank my parents, Srinivasan and Poongothai, for their endless love, support and confidence. They have given me the freedom to pursue my ambitions. I am who I am because of my parents and for that, I am blessed. I thank my younger brother, Rohit, for his prayers, support, and friendship. I would like to thank my in-laws, Rajendran and Chamundeswari, and grandmother, Vasanthalakshmi, for their support and prayers. I thank Bhuvana madam and Emeline madam for making me feel at home while away from home. Finally, I would like to thank all my extended family members for their love and affection.

Dedicated to my wife, Suchithra, who has stood with me through thick and thin

Chapter 1

Introduction and Motivation

Planning and effective utilization of resources plays an important role in improving any business. Scheduling helps in determining the order and time at which an entity arrives into the system such that the performance measures (e.g., utilization, waiting time, due date) are optimized.

1.1 Importance of Hospital Appointment System

Healthcare spending accounts for 17.4% of US GDP in 2013. The total spending in healthcare reached to $2.9 trillion, or $9,255 per person (National Health Expenditures 2013, 2014). Moreover, for the next 10 years, healthcare spending is expected to grow 1.1% faster than the US GDP and the contribution of healthcare spending to US GDP is expected to rise steadily as shown in Figure 1.1 (National Health Expenditures Projection 2014-2024, 2015).

One of the main reasons for the increase in healthcare spending is due to the increase in the number of patients who visit the hospitals (i.e., increase in demand). The number of patients visiting the hospital increases due to various reasons such as aging population and increase in number of insured customers. As a result of the Affordable Care Act (ACA), the percentage of US population without health insurance has dropped by 8.8 million in the year 2014 (the highest since 2008). The number of Americans without health insurance

was 33 million in 2014, down 21% from 2013 (Oaklander, 2015)

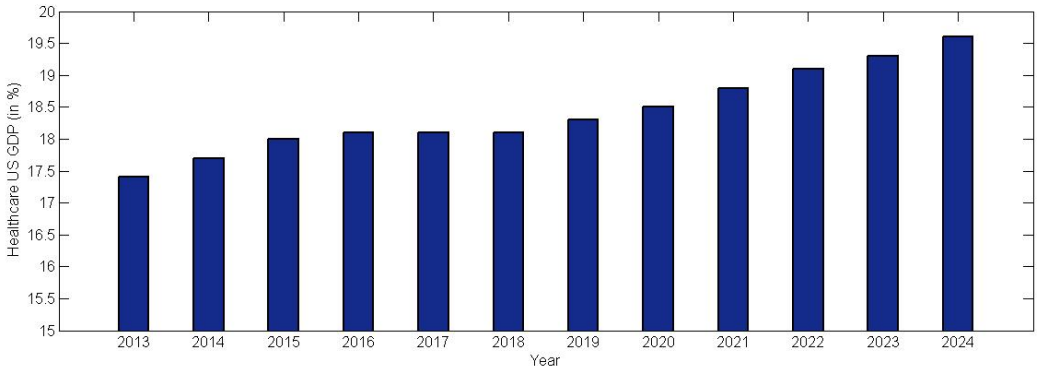


Figure 1.1: Healthcare spending as a % of US GDP

Hospitals employ one of their expensive resources, namely, doctors/physicians to serve the increasing patient demand. According to a report in TIME magazine (Oaklander, 2015), the US is expected to have a shortage of 90,000 physicians by 2025. It is evident that the demand for health care is expected to increase, while the supply of physicians to provide the care is expected to decrease. Therefore, fewer resources are available to meet the increasing demand and there is a growing pressure for the hospitals and clinics to improve the physicians' productivity/efficiency. Moreover, timely access to healthcare is essential to achieve good health outcomes and patient satisfaction.

Considering all the above factors, optimal use of the physician time is extremely important to improve the quality of care. An appointment system aims to distribute the workload (demand) throughout the operating hours of the clinic to ensure effective physician utilization and to improve patient satisfaction by providing timely access. Therefore, hospital appointment system would play a key role in managing the increasing patient demand in the future.

1.2 Scheduling in Healthcare Industry

In recent years, the healthcare industry has striven to improve its performance and become more efficient. The main inputs to a healthcare system are its patients and they can be

broadly classified as inpatients, outpatients, and emergency patients (Kopach et al., 2007) as shown in Figure 1.2. Patients who stay in the hospital overnight for treatment are called inpatients, while those who can be treated without being admitted to the hospital are called outpatients. On the other hand, patients who arrive to the hospitals for emergency situations are called emergency patients. Resources at the hospitals such as doctors/physicians, labs, and equipment, must be utilized efficiently to provide service to the patients at a faster rate and to achieve patient satisfaction.

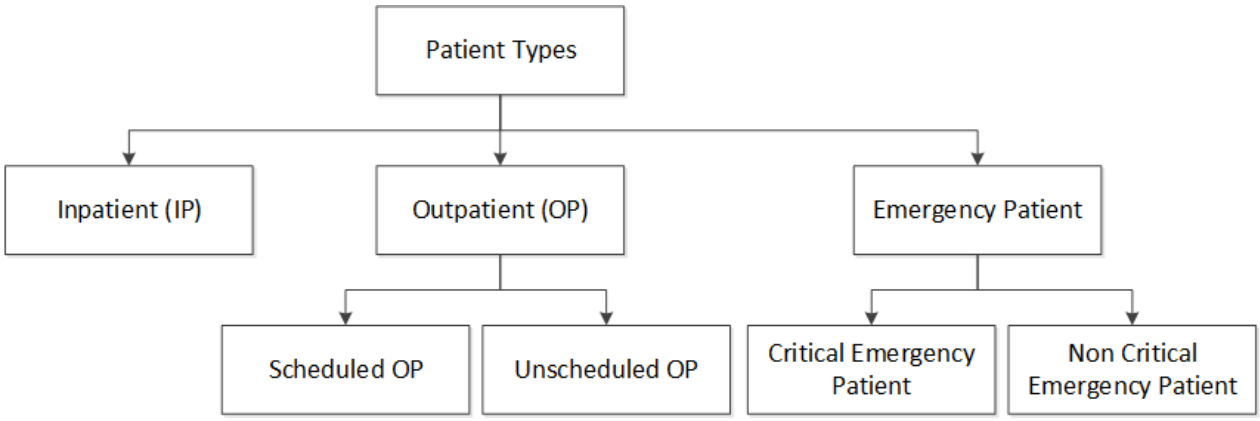


Figure 1.2: Classification of patients

A healthcare appointment system is used by the hospitals to provide patients access to their resources by specifying the time during which a patient can be served. The primary objective of using an appointment system is to identify the best ways to schedule the patients such that the (i) patients are satisfied (can be measured in terms of patient waiting time) and the (ii) resources are efficiently utilized (can be measured in terms of resource overtime and idle time).

The hospital's operating hours is generally divided into smaller time duration called *slots*, and the duration of the slots may be constant or varying. A patient is scheduled to one of the slots during which the patient undergoes treatment using the hospital's resources, such as physician, nurse, equipment. A well designed scheduling system has the potential to improve resource utilization and patient satisfaction. Understanding the importance of scheduling, hospitals have adopted different appointment types to achieve a balance

between resource utilization and patient satisfaction.

1.2.1 Appointment Types

Different appointment types and techniques have either been adapted from manufacturing and airline industries, or have been developed based on the characteristics of the healthcare industry (Pinedo, 2008). The following represent some of the major methods of outpatient scheduling in healthcare (Lindh et al., 2010):

- *Traditional Scheduling/Pre-booking (PB)*: The patients are scheduled to a future date weeks/months in advance. Therefore, the schedule is completely booked in advance allowing the hospitals to allocate its resources appropriately. However, the long time gap between the patient's call for appointment and the actual appointment date leads to uncertainties like patient no-shows and could make the schedule inefficient.

- *Walk-ins (WI)*: Walk-ins are very common due to their ability to maintain a steady flow of patients. The patients are served on a first come first serve basis. The utilization of resources and patient waiting time is highly uncertain as it depends on the number of patients visiting each day.

- *Open Access (OA)*: Specific slots or time periods in a schedule are left open to accommodate the patients on the same day or within the next 72 hours of receiving a request for an appointment, rather than scheduling the patient months later. This method is very similar to the Just in Time (JIT) system applied to the manufacturing sector in which the work is completed when and as it arrives. In OA system, appointment delay is reduced (i.e., time between patients call for appointment and the actual appointment date).

- *Double-booking*: In double-booking, two patients are booked for the same slot in order to compensate for the patient no-shows and to increase the utilization of resources. However, if both patients show-up for a particular slot, then the waiting

times of the remaining patients (i.e., the scheduled patient, who is treated second in that slot and those patients, who are expected to be treated in the upcoming slots) in the schedule increase and the physician has to work overtime to serve all the patients. Hence, careful planning and sequencing must be used while double-booking.

- *Wave Scheduling*: In this technique, many patients are scheduled every half hour and are served on first come first serve basis. The goal is to have the patients arrive in "waves" so that there are sufficient patients throughout the operating hours to ensure resource utilization. For example, if four patients are given a 10:00 AM appointment, then they will be treated as and when they arrive. This is commonly used in facilities that have multiple resources.

1.2.2 Appointment Rules

Appointment rule is a guideline developed based on clinic characteristics (e.g., average service time of patients visiting the clinic) to determine the appointment duration of each slot, the number of patients to be scheduled in each slot and list of slots to which a patient can be scheduled. The choice of the appointment type dictates the day when a patient can visit the clinic. For example, if a clinic practices open access appointment type, then a patient can visit the clinic on the same day. However, if the clinic practices pre-booking appointment type, then a patient may have to wait for days or weeks to visit the clinic. The appointment rule specifies the slot (i.e., time and duration) during which a patient is expected to be served and therefore, is necessary to implement the appointment system. Most of the appointment rules aim to reduce the patient waiting time, physician overtime and idle time. The most commonly used appointment rules are as follows:

- *Individual Block Fixed Interval (IBFI) Rule*: Patients are scheduled to individual blocks referred to as slots and the time interval between two patients is equal to the mean service times of patients. In other words, each patient is scheduled to a slot and the slot duration is constant throughout the clinic session.

- *2ATBEG*: Two patients are scheduled at the beginning of the clinic session and the remaining patients are scheduled at fixed intervals equal to the average service time of the patients.
- *OFFSET Rule*: The appointment duration for slots earlier in the clinic sessions are shorter than the mean service times and the appointment duration for slots later in the clinic session are longer than the mean service times of patients.
- *LVBEG Rule*: Patients with low service time variation are scheduled in the beginning of the clinic session while the remaining patients are scheduled at the end of the clinic session.
- *DOME Rule*: The appointment duration gradually increases in the middle of the clinic session and then decreases at the end of the clinic session resulting in a dome pattern.

The appointment types and the appointment rules together define an appointment system of a hospital or clinic. For example, an outpatient department might only accept patient calls for appointment weeks/months in advance (pre-booking method) and schedule one patient in each slot with a duration equal to the mean service time (IBFI rule). The best appointment system for a health system can be established by using historical data, such as the number of patient calls for appointment, service time of the patient, patient no-show rate, etc. The historical data serve as inputs to schedule patients using different combinations of appointment types and appointment rules. The different combinations are then evaluated using computer simulation to determine the best appointment system (i.e., an appointment system which improves the performance measures namely patient waiting time, number of patients denied appointment, resource overtime and resource idle time).

1.2.3 Importance of Outpatient Departments

Healthcare cost in the US is projected to rise at a rate of 6% per year (National Health Expenditure Projections, 2014). The projection also indicates that the average cost of

outpatient service has increased by 40% in the last five years. However, the increase in cost does not necessarily transform into better service to the patients. According to a survey by Hawkins (2009), the average appointment delay (time between the patient's call for appointment and the actual appointment date) is more than 2 weeks and may go up to 45 days. Moreover, on the day of the patient's visit, the average patient waiting time (total length of time spent by the patient waiting to be served/treated during their visit) is around 23 minutes. This is mainly because 90% of patient care in the US is provided by approximately 200,000 outpatient departments or outpatient clinics (Hawkins, 2009). Therefore, improving outpatient department's operations could positively impact the US health system. This has led to significant focus on improving the appointment system of outpatient departments in the literature (LaGanga and Lawrence, 2012). The flow chart shown in Figure 1.3 illustrates the various steps associated in scheduling a patient. Realizing the importance and scope associated with the outpatient departments, hospital administrators emphasize the importance of improving the patients' experience by reducing their appointment delay as well as the waiting time during their visit. Mogha et al. (2014) conducted a study to estimate the technical efficiency of 50 hospitals and found that, on average, hospitals have to increase 20% of their outputs with existing level of inputs to be efficient.

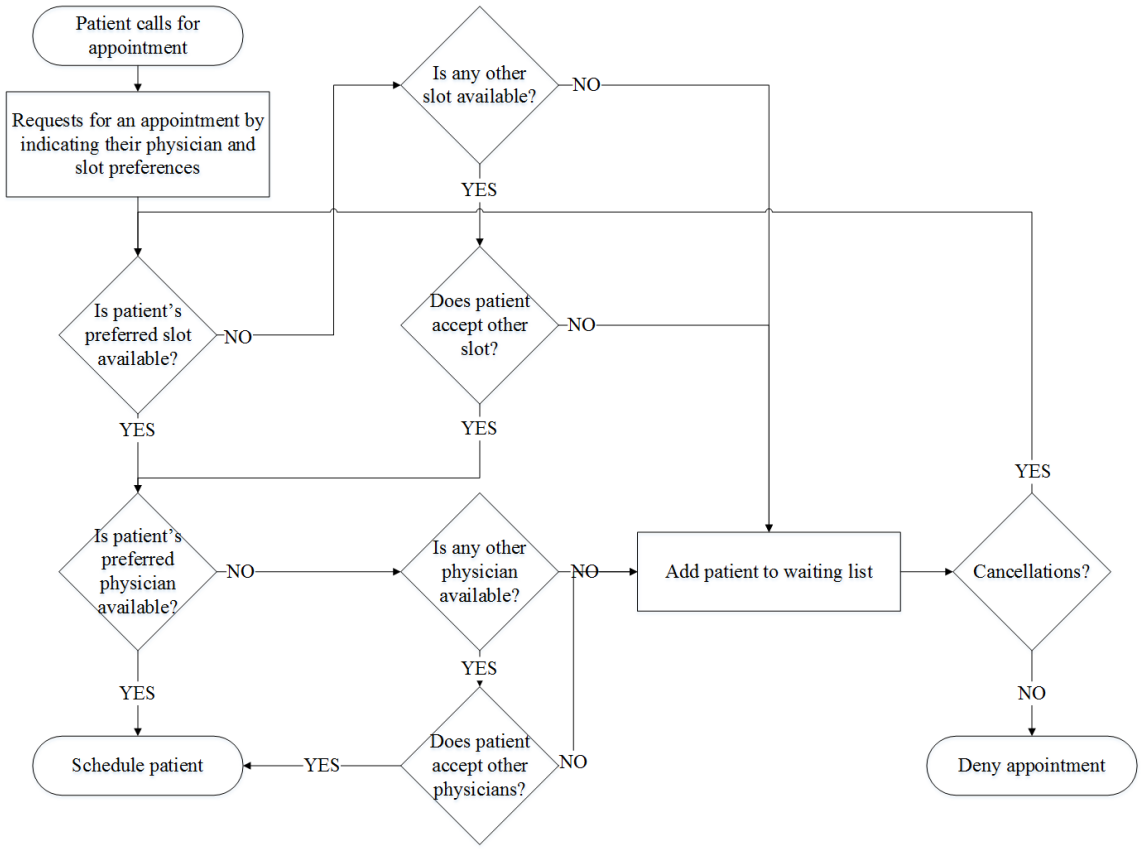


Figure 1.3: Overview of patient scheduling process

1.2.4 Nature of Patient Flow in Outpatient Departments

The scheduled patients are expected to arrive at the outpatient department on the day of their appointment and move through a sequence of phases. The movement of patients through a sequence of processes/phases is termed as patient flow. Generally, the nature of patient flow involves multiple phases. For example, an arriving patient may move through a sequence of stations/phases, such as check-in, visit nurse, visit physician and check-out as shown in Figure 1.4. The number of phases and the number of resources at each phase depends on various factors, such as the type of services provided by the outpatient department, medical history required for the services provided, billing procedure adopted by the hospital administration, and the number of patients visiting the clinic. In this particular example, all the patients must complete the check-in process and visit with one of the two available

nurses. Then the patients must visit their assigned physician (one of the four available physicians) for treatment and then check-out of the system.

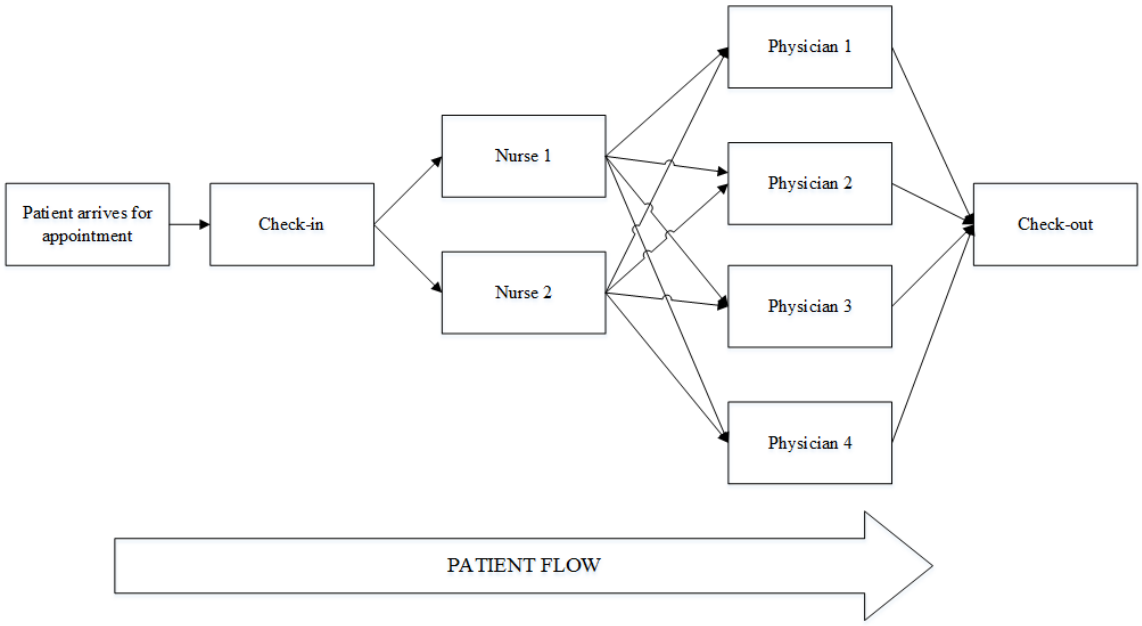


Figure 1.4: Multi-phase nature of patient flow

1.2.5 Challenges Associated with Outpatient Scheduling

The planning and scheduling of patients and resources in outpatient departments are challenging. The healthcare managers must anticipate uncertainties at each phase of the scheduling process (i.e., from patient's call for an appointment to the time when the patient completes the treatment on the day of the visit) to ensure patient satisfaction, resource utilization and low operating costs.

- *Uncertainty associated with patient calls*: The first step in scheduling an appointment is initiated by the patient when they call for an appointment. The number of patient calls for appointment for a given day and the time at which they call for appointment are uncertain and difficult to predict. If an outpatient department experiences a high number of patient calls for a given day, then it must deny appointments for certain patients on that particular day and must schedule the patient for a later date. This

might lead to patient dissatisfaction. On the other hand, if the number of patient calls for a given day is less, then the resources will be idle during clinic operating hours. The idle time of the resources for a given day cannot be inventoried and hence the outpatient department loses revenue for that particular day. Therefore, the outpatient department must take into account the variation in number of patient calls for a given day to balance patient satisfaction and resource utilization.

- *Patient preferences and availability*: Patients might prefer to meet with a specific provider (physician) in the outpatient department based on their prior experiences. In addition, the availability of each patient to visit the outpatient department might vary. The total number of patient calls for appointment for a given day varies and are divided among the physicians in the outpatient department. If many patients prefer the same physician then a particular physician is utilized more while the remaining physicians are utilized less. Therefore, to ensure patient satisfaction, it is important to design the schedule considering patient preferences and availability and to ensure effective physician utilization.

- *Patient no-show and late cancellation*: When a patient, scheduled to arrive at a given time, misses the appointment or when a patient cancels the appointment at the last moment, then the resources would remain idle during the clinic session. Therefore, patient no-shows and late cancellations must be taken into consideration when scheduling patients.

- *Service time variation*: The service time for each patient varies due to several reasons, such as the patient age, patient type (i.e., the type of treatment required for the patient), physician workload, time of the day. The variation in service time affects the patient's service completion time. As a result, other patients may have to wait before beginning their service. Hence, the variation in service time impacts the patient waiting time and therefore, must be handled while designing the appointment schedule.

1.2.6 Need for a New Appointment System

Patient's waiting time, appointment delay and the cost of treatment are some of the key measures that impact patient satisfaction (Naidu, 2009). Moreover, an outpatient department with longer appointment delay is expected to experience higher patient no-shows (i.e., a patient misses their scheduled appointment without informing the hospital). Patient no-shows negatively impacts the hospital revenue and the resource utilization. Hence, it is essential to use an appointment system that reduces patient waiting time, appointment delay and the cost of the treatment, while ensuring that the resources (e.g., physician, equipment) are neither over-utilized nor under-utilized.

Generally, one of the appointment types listed in Section 1.2.1 is used as a basis to design the appointment system of an outpatient department. A comparison of the pre-booking (PB) and open access (OA) methods is shown in Table 1.1. The PB method is more physician-centric because it aims to minimize physician idle time by scheduling many patients ahead of time. However, long appointment delay in PB leads to patient dissatisfaction and higher no-show rates. The OA system is best suited for outpatient departments experiencing high no-show rates. Even though OA has proven to reduce the patient no-show rate and appointment delay, it wouldn't be effective if the healthcare facility does not match the daily provider capacity to the patient calls for appointment (demand). OA and walk-ins (WI) are patient-centric approach because they reduce the appointment delay significantly. However, it could increase physician idle time because OA or WI demands are not known until they arrive.

Table 1.1: Comparison of open access and pre-book characteristics

Characteristics	**Open Access**	**Pre-Book**
Patient no-show rate	Low	High
Appointment delay	Low	High
Time to third appointment	Earlier	Delayed
Provider workload	Fluctuates	Balanced
Implementation	Difficult	Easy

Hence, each appointment type tries to maximize a performance measure at the ex-

pense of another performance measure. Therefore, it might not be possible to achieve a balance between resource utilization and patient satisfaction when using one particular appointment type. Hence, a Hybrid Appointment System (HAS), a combination of two or more appointment types, can be used to meet the needs of the patient and the outpatient department/clinic. OA reduces the appointment delay, while PB provides steady patient flow. Thus, in a hybrid appointment system, some clinic sessions are reserved weeks in advance, while the remaining clinical sessions are left open for same day appointments. A schematic representation of the HAS is shown in Figure 1.5.

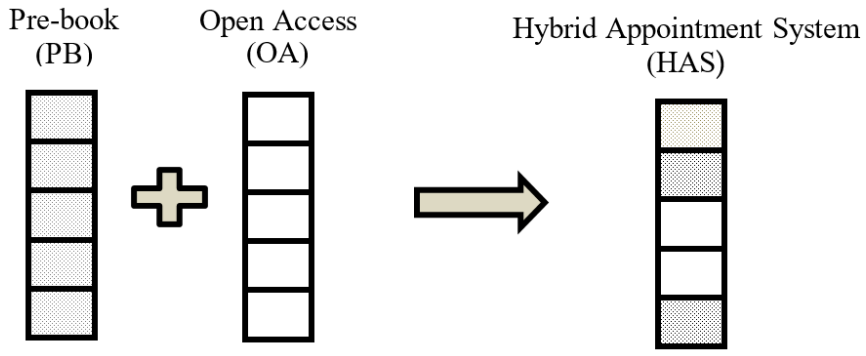


Figure 1.5: Schematic representation of a hybrid appointment system

1.3 Motivation

The appointment system must be robustly designed to achieve patient satisfaction and efficient resource utilization. Therefore, the following serve as a motivation for the proposed thesis:

1. The Affordable Care Act has increased the number of Americans with health insurance. This, in turn, has increased the patient demand to see physicians. However, US is expected to have a shortage of 90,000 physicians by 2025. Thus, a good appointment system design is essential to handle the increasing demand with limited capacity.

2. The rising healthcare costs could be controlled by improving the efficiency of out-

patient department because 90% of the patient care is provided by approximately 200,000 outpatient departments.

3. The average waiting time of patients in an outpatient clinic is 23 minutes. The long waiting time of patients is one of the main reasons for patient no-shows and cancellations.

4. The average appointment delay (time between patient call for appointment and the actual appointment date) is 18 days. Some of the reasons for Emergency Department (ED) crowding is the long appointment delay of outpatient departments and uninsured patients. Further, a substantial number of patients visiting the ED are not in a state of emergency and can be treated at outpatient departments. As a result of ED crowding, the quality and access to health care is affected. Therefore, reducing the appointment delay for outpatient departments could encourage patients to visit outpatient departments, thereby reducing ED crowding.

5. Each appointment type aims to improve one performance measure at the expense of another performance measure. Therefore, there is a need for a new appointment system that combines two or more appointment types, to meet the needs of the patient and the clinic.

6. Open access scheduling of patients has gained popularity over the last decade, due to its positive impact on patient throughput and patient satisfaction. However, the uncertainty in patient arrivals prevents some clinics from adopting it.

7. Various sources of uncertainties, such as patient no-shows, patient calls for appointment and service time, increase the complexity of designing a scheduling system for an outpatient department.

8. Services provided by the resources (e.g., physicians) cannot be inventoried for later use. Therefore, idle times of the resources are very expensive.

9. In recent times, there has been a shift in focus from inpatient services to outpatient services due to rising healthcare cost and shortage of hospital beds. Therefore, the demand for outpatient services will increase and an effective appointment system is needed to handle increasing demand without compromising on patient satisfaction.

10. Hospitals are expected to receive 65% of their revenue from outpatient departments.

11. The reputation of a hospital also depends on its outpatient department as more patients visit the outpatient department compared to the hospital's inpatient services.

12. Advancement in healthcare delivery systems and data analytics provides the scope to develop a data driven appointment system.

1.4 Research Plan

It is evident that outpatient departments are an important component of healthcare delivery system and improving their operations would improve the healthcare system significantly. However, the complexity and the uncertainties associated with outpatient scheduling poses a major challenge. The focus of this thesis is to propose new methods and models that would aid healthcare managers and hospital administrators to make the following decisions under uncertainty:

- Determining the duration of the clinic session reserved for each appointment type and their respective position in the schedule. In other words, the hospital administration has to decide the number of slots reserved for each appointment type (e.g., open access and pre-book) and the position of each slot type in the schedule (e.g., whether the first slot is a PB slot or an OA slot).

- Identifying the best appointment rule that suits the clinic characteristics to schedule patients as and when they call for appointments.

- Handling conflicting criteria and determining the best compromise solution.

- Predicting whether a patient will show up for an appointment and using the predicted information to schedule a patient.
- Determining the slots to overbook and the slots to single book.
- Determining whether to incorporate the nature of patient flow (e.g., check-in, see nurse, see physician, check-out) while designing an appointment system.

1.5 Outline of Thesis

The thesis is organized as follows. A detailed review of the various scheduling techniques and the appointment rules in the literature will be presented and the gaps in the existing research will be identified in Chapter 2. Most outpatient department's nature of patient flow involves multiple phases (e.g., check-in see nurse, see doctor, check-out), while most of the recommendations in the literature consider a simplified system (i.e., single phase system). Moreover, each patient may be available only during certain time slots to visit the clinic. It is not evident if the recommendations suggested for a simplified system can be adapted by the multi-phase system without affecting the schedule outcomes (waiting time, resource idle time, resource overtime) significantly. Chapter 3 evaluates the impact of the nature of patient flow and patient availability on the schedule outcomes using simulation.

Chapter 4 focuses on designing a hybrid appointment system (combination of pre-booking and open access) for a multi-phase multi-provider setting by considering all the uncertainties in the scheduling environment. A deterministic mixed integer linear programming (MILP) model is proposed to determine the optimal percentage of appointments reserved for pre-booking and open access, and a scenario-based Monte Carlo approach is used to account for uncertainty. Finally, the best configuration for the hybrid appointment system is determined using heuristics. A case study with real data from a Family Medicine clinic in Pennsylvania is used to show the feasibility of the proposed approach.

The approach proposed in Chapter 4 can be used by the hospital administration to design a hybrid appointment system (e.g., the slots reserved for each appointment type and

the slots to be overbooked). In order to schedule the patients as and when they call, it is necessary to identify the appointment or scheduling rule (discussed in Section 1.2.2) that best suits the clinic and patient characteristics. Chapter 5 proposes a data-driven scheduling framework for sequential scheduling of patients as they call for appointments. The proposed framework uses patient-related data from various sources to develop a predictive model for no-shows, and then, leverages that information to develop several scheduling rules to efficiently schedule patients as and when they call for appointments.

Chapter 6 illustrates the application of the proposed scheduling framework using a case study with real data. The performance of the proposed scheduling rules are compared with the current practice for effectiveness and efficiency. Also, the key findings of the study and its usefulness to healthcare practitioners are discussed in detail. Chapter 7 provides a summary of this dissertation, highlights the key theoretical contributions, methodological contributions and practical implications of this research, and discusses the potential scope for future research.

Chapter 2

Literature Review

Research for increasing the efficiency of hospital appointment system has been carried out since the very early initiatives of Bailey (1952) and Lindley (1952). In the last 60 years, there has been a lot of research that focuses on scheduling patients to an outpatient department or clinic (Cayirli and Veral, 2003; Gupta and Denton, 2008). Cayirli and Veral (2003) provided a comprehensive review of the factors considered in the literature while modeling an outpatient clinic and designing an appointment system. The authors discussed the complications involved at each phase of the patient flow process and outlined the methods used in the literature to model the nature of patient flow at an outpatient clinic. In addition, they briefly discussed the various appointment rules that can be used in an appointment system. Gupta and Denton (2008) provided a comprehensive review of the different appointment system environments, such as primary care, specialty clinic and their challenges.

In this chapter, we will discuss the various patient scheduling models and appointment rules proposed in the literature. Table 2.1 shows the key features of the appointment systems proposed in the literature and is used for categorizing the literature in this chapter.

Table 2.1: Features of appointment system

Dimension	Attribute	Description
Appointment system characteristics	Appointment rules	Refers to the appointment rules studied or proposed
	Appointment type	Refers to the appointment types considered to design the appointment system
Outpatient department characteristics	Nature of patient flow	Refers to the number of stations (phases) that a patient must visit before leaving the system (hospital or clinic)
	Patient's provider preference	Denotes whether the scheduling model considers the patient's preference to be treated by a particular provider
	Patient availability	Denotes whether the scheduling model considers the patient's availability while scheduling
Model characteristics	Objective function	Refers to the goal of the proposed model
	Solution methodology	Refers to the methodology adopted to achieve the objective
	Performance measures	Refers to the key metrics that are used to evaluate different schedules
Environmental characteristics	Patient no-show	Specifies whether the no-show rate is constant for all patients or varies by patient (i.e., patient specific no-show)
	Service time	Refers to the nature of the service time considered in the model. It can be constant or a random variable
	Demand (Patient calls)	Refers to the nature of patient calls for appointments. It is either known (deterministic) in advance or random (stochastic)
	Patient lateness	Refers to the uncertainty in the arrival of patients to the clinics

2.1 Design of Appointment Systems

Initially, most research work tried to optimize and improve the pre-booking (PB) method by evaluating various appointment rules and determining the best setting that improves the performance measures (Brahimi and Worthington, 1991; Ho and Lau, 1992; Klassen and Rohleder, 1996).

Ho and Lau (1992) evaluated the various appointment rules for pre-booking method with the objective of minimizing the total cost (expected weighted sum of physician idle time cost and patients waiting time cost). The authors identified that a single best appointment rule does not exist and therefore, an appointment rule that minimizes the total cost for one particular clinical setting might not provide the least cost solution for another clinical setting. The authors concluded that three factors, namely patient no-show rate, coefficient of variation in service times and the number of patients scheduled, significantly affected the total cost.

Currently, increased competition and patients' higher expectations have resulted in significant modifications to the appointment system in hospitals and clinics. Many researchers have recommended that walk-ins (WI) must be expected and planned, while designing an appointment system by observing the pattern of patient arrivals (Fetter and Thompson, 1966; Shonick and Klein, 1977; Field, 1980). This type of arrival is most common in primary care clinics, where the patients are provided service without any pre-determined time slots or appointments (Barron, 1980). The clinics that are responsible for the patients' total care should consider including walk-ins in order to provide more efficient care. Therefore, hospital administrators have to forecast the number of patients who are expected to walk-in. Also, Taylor (1984) and Virji (1990) indicated that there is a tendency among lower socio-economic classes to visit the clinic without an appointment. These findings suggest that a clinic that denies access to walk-ins may further disadvantage these groups. Hence, in general, it is better to anticipate and include walk-in slots in an appointment system to avoid delays and increase patient satisfaction. Cayirli and Günes (2013) studied

the impact of adjusting capacity through intra-week seasonality of walk-ins and evaluated different appointment rules to determine the best design for an appointment system with seasonal walk-ins. Using a simulation optimization approach, they concluded that the performance measures of the clinic (patient wait time, physician idle time, and overtime) could be improved by accounting for seasonality in walk-ins compared to no seasonal adjustment. Further, the best position (slot) to overbook depended on the cost ratios of the three clinic performance measures.

In the early 1990s, open access (OA) scheduling system was introduced and their successful implementation resulted in improved patient satisfaction and resource utilization (Boelke et al., 1999; Herriott, 1999; Murray and Tantau, 2000; Kennedy and Hsu, 2003; Murray and Berwick, 2003; O'Hare and Corlett, 2004; Robinson and Chen, 2010). Some researchers concluded that OA policy was better than PB, in terms of reduced patient waiting time and no-show rates (Boelke et al., 1999; Murray and Tantau, 2000; Robinson and Chen, 2010). Sceptics of the OA system indicate that OA is difficult to implement and might reduce continuity of care (Murray and Tantau, 2000; Singer and Regenstein, 2003; Ahluwalia and Offredy, 2005). Patrick (2012) has indicated that if the cost of the appointment delay is high, then OA will be the optimal outpatient scheduling policy. However, when the appointment delay cost is low, then a policy with a small scheduling horizon performs better than OA.

Recently, there has been a significant focus on using a combination of several appointment types (Kopach et al., 2007; Qu et al., 2007; Lee and Yih, 2010; Robinson and Chen, 2010; Qu and Shi, 2011; Peng et al., 2014; El-Sharo et al., 2015). Studies that consider combining two or more appointment types have addressed several challenges regarding the configuration or design of the appointment system.

Kopach et al. (2007) studied the effects of continuity of care and clinic throughput on OA implementation. A simulation analysis was conducted, and the study suggested that transferring to OA would serve more patients in the long run without affecting the system. However, a quick transition to OA would lead to a decrease in the continuity of care of the

patients.

Muthuram and Lawley (2008) considered open access and pre-booking methods and proposed a stochastic overbooking model to determine the appointment time for each patient for a given day. Their model schedules patients in a sequential fashion (i.e., the schedule is built incrementally patient-by-patient) with the objective of improving patient waiting time, resource overtime and revenue. The authors showed that the objective function was unimodal and provided an optimal stopping rule beyond which the patients were rejected.

Liu et al. (2010) considered open access and pre-booking methods to determine the day on which a patient could be scheduled such that the long run net reward for the clinic was maximized. In addition, the authors considered patient no-shows and cancellations in their model. They compared their proposed heuristic with existing methods and concluded that the proposed heuristic performed well, especially when the demand was high. The authors also suggested that the combination of open access and pre-booking was better when the demand was low. However, the proposed heuristic did not specify the appointment time for each patient. Therefore, their model had to be integrated with the existing slot scheduling algorithm (e.g., Muthuraman and Lawley, 2008) so that the appointment date and time could be specified. El-Sharo et al. (2015) developed an overbooking model to maximize the profit for a multi-provider outpatient clinic using pre-booking and walk-ins. The authors concluded that the profits could be maximized by reallocating the overflowing patients (i.e., patient's whose service start time was delayed and extended to the next appointment slot) based on provider's workload.

The appointment rules in the literature have also evolved based on patient's and hospital's needs. Most appointment rules are designed by using two main parameters, namely block size and appointment interval. Block size refers to the total number of patients scheduled in a slot. A schedule that has exactly one patient in each slot has a block size of one (individual block). However, if a schedule has more than one patient in each slot and if the number of patients in each slots are same, then it is a multiple-block system. On the other hand, if a schedule has more than one patient in each slot and if the number of

patients in each slot varies, then it is a variable-block system. Appointment interval refers the duration of each slot and can be constant throughout the schedule (fixed interval) or can be varying (variable interval).

One of the earliest appointment rules, known as the Bailey's rule or 2ATBEG rule, was based on the assumption that the service time distribution of all the patients were the same (Bailey, 1952). The rule scheduled two patients at the beginning of the clinic session and the remaining patients at fixed intervals equal to the average service time of the patients.

The individual block fixed interval (IBFI) rule schedules one patient at equal intervals and has been studied for the last five decades (Fetter and Thompson, 1966; Klassen and Rohleder, 1996; Rohleder and Klassen, 2000). Similarly, many other rules, such as the multiple block fixed interval (MBFI), variable block fixed interval (VBFI) have been proposed and studied in the literature (Soriano, 1966; Rising et al., 1973; Fries and Marathe, 1981; Cox et al., 1986). These rules use fixed interval and the average service time of the clinic is primarily used as the interval length or appointment duration. Ho et al. (1995) performed an extensive study by proposing different variable-interval appointment (VI) rules (e.g., IBVI - individual block variable interval, MBVI - multiple block variable interval) that minimized patient waiting time, without significantly increasing the clinic idle time. Their rules had an appointment duration less than the average service time until first K patients were scheduled and then the appointment duration were greater than the average service time for the remaining patients. However, these appointment rules may not be very effective when some patients walk-in or when there are requests for same day appointments.

Klassen and Rohleder (1996) overcame this issue by considering same day appointments and were the first to evaluate different appointment rules in a dynamic environment (i.e. schedule patients as and when they call without knowing which types of patient will call at a later time). Their approach was more realistic for an actual clinic. They used the IBFI rule and proposed different special cases based on the fact that the scheduling manager would have knowledge about the patient's service time characteristics (i.e., low variance or

high variance in service time) when they called for appointments. The different rules were: low variance patients in the beginning of the clinic session (LVBEG), high variance patients in the beginning of the clinic session (HVBEG), low variance patients in the beginning and end of the clinic session (LVBND), high variance patients in the beginning and end of the clinic session (HVBND) alternating low and high variance patients (ALT1-1). They found that scheduling low variance patients in the beginning of the clinic session (LVBEG rule) performed well in most clinic settings. Recently, Cayirli et al. (2006) proposed a special case of IBVI rule called the DOME rule. In the DOME rule, the appointment duration increases initially and then decreases at the end of the day resulting in a dome pattern. The authors also proposed new rules by combining existing rules with the DOME rule: 2BGDM (combination of DOME and 2ATBEG) and MBDM (combination of DOME and MBFI). According to the authors, these rules performed well in the presence of no-show and walk-ins.

Table 2.2 shows the appointment system design characteristics of various outpatient scheduling models proposed in the literature by specifying the appointment type and the appointment rules used to develop the scheduling model.

Table 2.2: Appointment system design characteristics of outpatient scheduling

Author (Year)	**Appointment system design characteristics**	
	Scheduling method	**Appointment rules**
Bailey (1952)	PB	2ATBEG
Ho and Lau (1995)	PB	VI rules
Klassen and Rohleder (1996)	PB and OA	LVBEG, HVBEG, LVBND, HVBND, ALT1-1
Cayirli and Veral (2006)	PB and WI	DOME, 2BGDM, MBDM
Muthuraman and Lawley (2008)	PB and OA	VBFI
Glowacka et al. (2009)	*	Single block, IBFI, MBFI
Chakraborty et al. (2010)	*	IBFI, MBFI, DOME, MBDM, 2ATBEG
Daggy et al. (2010)	*	VBFI
Lee and Yih (2010)	PB and OA	VBFI
Liu et al. (2010)	PB and OA	†
Gul et al. (2011)	PB	IBVI
Qu et al. (2011)	PB and OA	IBVI
Hahn-Goldberg et al. (2012)	PB	IBVI
Patrick (2012)	OA and PB	VBFI
Saremi et al. (2013)	*	IBVI
Erdogan and Denton (2013)	PB and OA	VBVI
Peng et al. (2014)	PB, OA and WI	VBFI
El-sharo et al. (2015)	PB and WI	VBFI
Erdogan et al. (2015)	PB and OA	IBVI
Yan et al. (2015)	PB	MBFI

*The authors did not discuss about the appointment types considered in their paper.

† The authors did not discuss about the appointment rules in their paper.

2.2 Outpatient Department (System) Characteristics

Research that focuses on designing an appointment system determines the appropriate number of slots for each appointment type such that the performance measures are optimized. These studies consider simplified clinic scenarios by assuming the nature of patient flow to be single-phase (one station). In general, healthcare system involves the flow of patients through multiple-phases (Erdogan et al., 2015). Research that focuses on modeling the nature of patient flow uses process improvement techniques to optimize the performance measures and consider the actual nature of patient flow. However, most of these studies do not focus on the design of the appointment system (i.e., they do not determine the number of slots for each appointment type) and are specific to a particular hospital.

Cote (1999) studied the impact of patient flow on the resource utilization using a simulation model. The author used various statistical analyses and found that utilization of the resources might not be directly correlated with patient waiting time. Chand et al. (2008) studied the patient flow process to identify the sources of variability and methods to improve patient waiting time and resource utilization. They identified the sources of variability at each phase and suggested appropriate recommendations, such as adding an extra resource, processing one patient record at a time instead of batch processing. Rohleder et al. (2011) identified the appropriate number of resources required at each phase to minimize the waiting time. Zhecheng et al. (2012) identified the causes for clinic overtime and patient waiting time by analyzing the clinic process. They recommended evenly distributed appointment sessions, serving patients on first appointment first serve basis and starting the session on the scheduled time. Kuljis et al. (2001) developed a model for outpatient clinic to analyze the impact of system operation policies on the patient waiting time. Their generic model was tested at 20 different clinics and found that the model aided in reducing the patient waiting time.

In general, a healthcare system has multiple providers working in parallel (El-sharo et al., 2015; Erdogan et al., 2015). In addition, a patient may prefer to be seen by a particular

physician for several reasons, such as the physician might be aware of the patient's history, the physician has a very good experience in treating certain types of diseases. Therefore, a patient must be scheduled to his/her preferred physician to improve the continuity of care and patient satisfaction. Hence, it is essential to consider the patient's provider preference, while designing the scheduling models.

Saremi et al. (2013) proposed three heuristic based optimization models to schedule patients to a multi-provider multi-phase operating room department. In their model, they considered resource compatibility (i.e., each patient type can be treated only by a particular surgeon type). It is very similar to modeling patient's provider preference, because the resource compatibility ensures that the patient cannot be served by other surgeon types and it ensures that the patient is not scheduled to a physician whom the patient does not prefer. In addition, the authors studied existing appointment rules in scheduling using their proposed models and concluded that the appointment rules that present a dome pattern minimized patient waiting times. Further, a patient might be able to visit the clinic only at specified times (e.g., Fridays only, any day between Noon and 1:00 PM). The patient's availability may vary depending upon various factors, such as occupation, age, weather. Therefore, to develop a patient-centric appointment system, it is essential to incorporate their availability in the scheduling models.

Chakraborty et al. (2010) proposed a sequential scheduling model by considering patient availability (i.e., slot preference) to make decisions regarding patient admission and appointment times. In addition to patient preference, Yan et al. (2015) considered service fairness in their sequential scheduling model to determine the optimal number of patients to schedule and their appointment times such that the profit was maximized. The authors concluded that considering patient availability might increase the number of patients served on a given day but might reduce the overall profit due to less flexibility. The models proposed by Chakraborty et al. (2010) and Yan et al. (2015) employ a myopic policy and therefore, do not consider patients who may call in the future. Table 2.3 illustrates the characteristics of various outpatient scheduling models proposed in the literature.

Table 2.3: Outpatient department characteristics of outpatient scheduling models

Author (Year)	Outpatient department characteristics		
	Nature of patient flow	**Patient's Physician Preference**	**Patient availability**
Bailey (1952)	Single-phase	Not applicable	Not considered
Ho and Lau (1995)	Single-phase	Not applicable	Not considered
Klassen and Rohleder (1996)	Single-phase	Not applicable	Not considered
Cayirli and Veral (2006)	Single-phase	Not applicable	Not considered
Muthuraman and Lawley (2008)	Single-phase	Not applicable	Considered
Glowacka et al. (2009)	Multi-phase	Not applicable	Not considered
Chakraborty et al. (2010)	Single-phase	Not applicable	Considered
Daggy et al. (2010)	Single phase	Not considered	Not considered
Lee and Yih (2010)	Single-phase	Not applicable	Not considered
Liu et al. (2010)	Single-phase	Not applicable	Not considered
Gul et al. (2011)	Multi-phase	Not applicable	Not considered
Qu et al. (2011)	Single-phase	Not applicable	Not considered
Hahn-Goldberg et al. (2012)	Multi-phase	Not considered	Not considered
Patrick (2012)	Single phase	Not applicable	Not considered
Saremi et al. (2013)	Multi-phase	Considered	Not considered
Erdogan and Denton (2013)	Single-phase	Not applicable	Not considered
Peng et al. (2014)	Single-phase	Not applicable	Not considered
El-sharo et al. (2015)	Single phase	Not considered	Not considered
Erdogan et al. (2015)	Single-phase	Not applicable	Not considered
Yan et al. (2015)	Single phase	Not applicable	Considered

The patient's physician preference is not applicable because it is a single server system

2.3 Environmental Characteristics of Outpatient Departments

There are various environmental factors that could disrupt an appointment schedule. In reality, the number of patients who call for an appointment for a given day is random. Moreover, the calls occur in a sequential manner and each patient must be given an appointment before the call ends. Hence, the scheduling manager must decide whether to schedule the patient or not without knowing the number of calls that could come later and the decision is based on the appointment system (appointment types + appointment rules) adopted by the hospital. Most research papers assume that the patient demand is known ahead and then identify the best sequence to schedule the patients. However, some papers considers the uncertainty in demand, while scheduling patients.

Hahn-Goldberg et al. (2012) addressed the uncertainty in demand by proposing a dynamic template scheduling that included open access and pre-booking methods. The schedule configuration/template (number of slots reserved for PB and OA) are updated as and when a demand occurs, using a deterministic optimization model, such that the makespan is minimized. The authors tested the proposed model at a chemotherapy center and identified that the makespan decreased by 20% compared to the existing appointment system.

Erdogan et al. (2015) proposed a stochastic integer programming model to schedule patients under demand uncertainty. The authors consider two categories of patients, namely routine patients and urgent patients. Their model determines the appointment times and the sequence such that the weighted sum of patient waiting time, physician overtime and physician idle time are minimized. The authors conclude that the optimal appointment schedule structure and the sequencing decisions are sensitive to the cost parameters and provide conditions when simple heuristics perform well.

The uncertainty in demand occurs even before the patient visits the outpatient department for service. There are other environmental characteristics that impact the schedule during the patient's visit to the outpatient department. In most cases, a patient's visit to

the outpatient department and the nature of patient's flow is modeled as a queuing system (e.g., arrival, service, departure). At both the arrival and service stages, there are several environmental characteristics that affect the schedule.

For arrival, the entire schedule is affected if the patients arrive late for their appointments. Even if a patient arrives a few minutes late to his/her appointment, the effects on the schedule might be magnified because the service might not begin as scheduled for the subsequent patients who might have arrived on-time. In addition, the physicians must work longer to serve all the patients. Therefore, late arrivals lead to patient dissatisfaction and overtime for the physicians. In some situations, early arrivals are also undesirable because early arrivals might lead to waiting room congestion leading to patient dissatisfaction (Welch and Bailey, 1952). In order to encourage the patients to arrive on-time, hospitals generally rewards a patient who has a history of arriving on-time by giving appointments at the dates requested by the patients (Barron, 1980). In some situations, the appointment system is designed to include some slack (Nolan et al., 2004) to reduce the effects of late arrivals and to compensate for unexpected arrivals (i.e., emergency cases or walk-ins). For instance, an outpatient department with an 8 hour shift may have 16 slots with each slot having a constant duration of 30 minutes. Instead, the appointment system may be designed to incorporate a slack slot after every six slots. In this case, there will be 14 slots in an 8-hour shift. The slack slot provided after six slots would help minimize the delay factor increasing the whole day and could increase patient satisfaction (Nolan et al., 2004).

The arrival situation becomes more complicated if the patient does not show up for the appointment. In situations like these, the hospital/clinic loses revenue (Ho and Lau, 1992; Murdock et al., 2002). Further, no-show impacts clinic productivity and provider efficiency (Nancy et al., 2012). It is impossible to eliminate no-shows but they can be reduced by adopting various strategies, such as telephone reminder and open access scheduling (Laiyemo et al., 2014). He et al. (2013) developed a simulation model for policy making and found that overbooking reduces appointment delays and compensates for no-shows. In addition, some studies predict the no-show rate and use it to schedule the patients

effectively.

Glowacka et al. (2009) determined the optimal number of patients to be scheduled in a clinic session by predicting the patient no-show value using association rule mining. However, their model assigned no-show values only to some patient groups. Daggy et al. (2010) overcame this issue by predicting patient-specific no-show values using the data collected from Veterans Affairs Medical Center. The authors developed a logistic regression model and scheduled patients based on their no-show outcomes, with the objective of minimizing the total cost (sum of waiting time, overtime and utilization). They compared their model with the current practice and concluded that scheduling models that considered individual patient no-show values might improve the schedule efficiency without limiting access to care for patients, who were likely to miss their appointments.

Samorani and LaGanga (2015) used data mining techniques to predict the no-show outcomes of new appointment requests, based on the past appointment data. They developed a heuristic to schedule the patients based on the predicted show outcome and concluded that same day appointments should be given for likely shows, while future day appointments be given for likely no-shows. The authors tested various configurations/clinic settings and observed that the patient waiting time, number of unscheduled patients and resource utilization improved in most cases.

For service, the time required to treat each patient is random and depends upon several factors that include: (i) experience of the doctor, (ii) availability of resources, (iii) complexity of the disease (Rising et al., 1973). Cayirli and Veral (2003) indicated that the coefficient of variation (CV) of service time for an outpatient clinic varies between 0.25 and 0.45. Therefore, it is complicated to determine the time for each slot for different types of patients since the entire process is stochastic (Rising et al., 1973). It was observed that a high variance in service times would negatively impact the performance of the doctors, and also would increase patient waiting time (Gupta and Denton, 2008). In order to address this issue, Kopach et al. (2007) conducted a study to determine the time slots for each patient type. They concluded that in practice an equally spaced time slot would perform the same

as the different time slots. Therefore, irrespective of the patient's type, the patients should be provided with equal time slots to compensate for the service time variation.

Table 2.4 shows the environmental characteristics of various outpatient scheduling models proposed in the literature. The characteristics include the nature of patient no-show (constant or predicted), demand (known or random), service time (deterministic or stochastic), and patient lateness.

Table 2.4: Environmental characteristics of outpatient scheduling models

Author (Year)	Environmental characteristics			
	Patient no-show	**Service time**	**Demand**	**Patient lateness**
Bailey (1952)	Not Applicable	Gamma distribution	Known	Not considered
Ho and Lau (1995)	Constant	Uniform and Exponential	Known	Not considered
Klassen and Rohleder (1996)	Constant	Log-normal distribution	Random	Considered
Cayirli and Veral (2006)	Constant	Log-normal distribution	Known	Considered
Muthuraman and Lawley (2008)	Constant	Exponential distribution	Random	Not considered
Glowacka et al. (2009)	Patient specific	Log-normal distribution	Known	Not considered
Chakraborty et al. (2010)	Constant	General distribution	Random	Not considered
Daggy et al. (2010)	Patient specific	Log-normal distribution	Random	Not considered
Lee and Yih (2010)	Patient specific*	Constant	Known	Not considered
Liu et al. (2010)	Patient specific*	Not applicable	Random	Not considered
Gul et al. (2011)	Constant	Log-normal distribution	Known	Not considered
Qu et al. (2011)	Constant	Not Applicable	Known	Not considered
Hahn-Goldberg et al. (2012)	Constant	Empirical distribution	Random	Not considered
Patrick (2012)	Patient specific*	Constant (deterministic)	Random	Not considered
Saremi et al. (2013)	Constant	Random	Known	Not considered
Erdogan and Denton (2013)	Constant	Random	Unknown for OA	Not considered
Peng et al. (2014)	Constant	Uniform distribution	Known	Not considered
El-sharo et al. (2015)	Constant	Constant	Known	Not considered
Erdogan et al. (2015)	Constant	Random	Random	Not considered
Yan et al. (2015)	Constant	Exponential distribution	Known	Not considered

*Depends only on appointment delay

2.4 Types of Outpatient Scheduling Models

Simulation has proven to be a very useful tool to model the healthcare environments and analyze the impact of different scenarios (Barnes, 1997). Lee and Yih (2010) performed a simulation study to analyze the effect of OA scheduling in outpatient clinics. They considered the environmental conditions to analyze the impact of OA scheduling by varying the proportion of slots for OA and PB. The results of the simulation study indicated that OA is beneficial, especially when the patient no-show rates are significant.

Recent research has focused on designing a hybrid appointment system by using analytical and heuristic approaches to determine the best configuration of the appointment system. Qu et al. (2007) proposed a closed form solution to determine the optimal number of OA appointments to match the daily demand. The authors concluded that the number of OA appointments depends on the OA demand, provider capacity and no-show rates. A more detailed design or configuration of a hybrid schedule was proposed by Peng et al. (2014) using a heuristic based approach. The authors specified the position of the OA appointments in the schedule and also the slots which must be double booked. They concluded that the optimal template was significantly affected by no-show rate, demand and cost coefficients.

A schedule must be evaluated for its effectiveness. There is a lot of teamwork and cost factors involved in determining a schedule, such as the cost of hiring people to perform data analysis and determining patient no-show probability (Cayirli and Veral, 2003). An ineffective schedule does not improve the efficiency of the hospital and leads to a decrease in their customer base. Therefore, various performance measures have been used to evaluate the effectiveness of a schedule. They can be broadly classified as cost related, time related and other measures (Cayirli and Veral, 2003). The measures involving cost assign a cost factor for all the sequences of events in the hospitals. A dollar amount is incurred for each activity, such as the number of patients in the waiting room, the number of workers, and the utilization of physicians. The time related measures include average time for a doctor

to serve a patient, average number of hours a doctor is idle and average number of hours a doctor performs overtime work. Other performance measures include the productivity of the physician, utilization of equipment, number of patients waiting to be served, and number of patients served. In general, the goal of all schedules is to provide high patient satisfaction, have full utilization of the resources, and be cost effective.

Most of the outpatient scheduling models in the literature combine multiple performance measures into a single objective by converting the value of the performance measure to an approximate dollar amount (Cayirli and Veral, 2003; Qu et al., 2007; Patrick, 2012; Yan et al., 2015). Thus, the objective function, in most cases is to minimize the total cost or to maximize net profit. However, some scheduling models in the literature (e.g., Lee and Yih, 2010; Gul et al., 2011) have considered multiple objectives and obtained a set of non-dominated solutions (i.e., a solution in which none of the objectives can be improved without worsening at least one of the other objective values).

Table 2.6 shows the characteristics of the various outpatient scheduling models proposed in the literature, by specifying the nature of objective function (single or multiple), solution methodology and the performance measures used to evaluate the schedule.

Table 2.5: Model characteristics of outpatient scheduling models

Author (Year)	Model characteristics		
	Objective function	**Solution methodology**	**Performance measures**
Bailey (1952)	Single	Mathematical model	• Patient waiting time • Physician idle time
Ho and Lau (1995)	Single	Simulation	• Patient waiting time • Physician idle time
Klassen and Rohleder (1996)	Single	Simulation	• Patient waiting time • Physician idle time
Cayirli and Veral (2006)	Multiple	Simulation	• Patient waiting time • Physician idle time • Physician overtime
Muthuraman and Lawley (2008)	Single	Simulation based optimization	• Patient waiting time • Physician overtime • Revenue
Glowacka et al. (2009)	Single	Association rule mining and Simulation	• Patient waiting time • Resource idle time • Resource overtime • Revenue
Chakraborty et al. (2010)	Single	Simulation based optimization	• Patient waiting time • Physician overtime
Daggy et al. (2010)	Not Applicable	Simulation	• Patient waiting time • No. of patients served • Physician utilization • Physician overtime
Lee and Yih (2010)	Multiple	Simulation	• Patient waiting time • No. of patients rejected • Clinic utilization
Liu et al. (2010)	Single	Heuristic	• Reward
Gul et al. (2011)	Multiple	Simulation and Genetic Algorithm	• Patient waiting time • Resource overtime

Table 2.6: (cont'd) Model characteristics of outpatient scheduling models

Author (Year)	Model characteristics		
	Objective function	**Solution methodology**	**Performance measures**
Qu et al. (2011)	Single	Deterministic Optimization	• Expected no. of patients served • Variance of the no. of patients served
Hahn-Goldberg et al. (2012)	Single	Deterministic optimization	• Makespan
Patrick (2012)	Single	Simulation and Dynamic programming	• Appointment delay • Physician idle time • Physician overtime • Revenue
Saremi et al. (2013)	Single	Simulation and Metaheuristics	• Patient waiting time • Patient completion time • No. of patients rejected
Erdogan and Denton (2013)	Single	Stochastic programming	• Patient waiting time • Physician overtime • Physician utilization
Peng et al. (2014)	Single	Simulation and Genetic Algorithm	• No. of patients served • Patient waiting time • Physician idle time • Physician overtime
El-sharo et al. (2015)	Single	Simulation	• Patient waiting time • Physician idle time
Erdogan et al. (2015)	Single	Stochastic programming	• Patient waiting time • Physician idle time • Physician overtime
Yan et al. (2015)	Single	Simulation	• Patient waiting time • Physician idle time • Physician overtime • Revenue

2.5 Observations from the Literature and Research Opportunities

In this section, some observations are presented based on the detailed review of literature on outpatient scheduling models. In the early 1950s, the design of an appointment system included the evaluation of various appointment rules for the traditional scheduling or pre-booking method. Research on improving the efficiency of outpatient scheduling started with the initiative of Bailey (1952), where the author proposed the 2ATBEG appointment rule (i.e., scheduling two patients at the beginning of the clinic session and the remaining patients at fixed intervals equal to the average service time of the patients). Since then, several researchers have proposed different appointment rules to improve the efficiency of outpatient scheduling models with respect to various performance measures (e.g., patient waiting time, profit, physician overtime). In the early 1990s, the open access (also known as advanced access) appointment type was introduced, where the patients are provided an appointment within 72 hours of the patient's call for appointment, thereby, reducing the appointment delay. In addition, the walk-in method have also gained significant recognition due to increased need for the hospitals to stay competitive and achieve patient satisfaction. Thus, many researchers focused on designing an open access appointment system and reported the advantages and disadvantages of implementing a pure open-access appointment system. In recent years, most of the outpatient scheduling models aim to design and improve the efficiency of a hybrid appointment system (i.e., combination of two or more appointment types).

2.5.1 Opportunities to Improve the Appointment System

The observations presented below are based on the review of literature on the appointment system design and outpatient department characteristics of the outpatient scheduling models (Sections 2.1 and 2.2).

- The appointment rules in the literature are mostly based on observed patterns or extension of existing rules. None of the appointment rules are data-driven. In other

words, the appointment rules do not use the available data (electronic medical records) to schedule patients.

- Almost all the outpatient scheduling models consider the nature of patient flow to be single-phase, while designing an appointment system. However, in reality most healthcare systems involve the flow of patients through multiple-phases (e.g., check-in, visit nurse, visit physician, check-out). Hence, the design of the appointment system for a single-phase clinic environment may not be suitable for a multi-phase environment.

- Most models consider a single-provider setting (one physician only) to design the appointment system. In addition, models that consider multiple-providers mostly ignore the patient's provider preferences, while designing an appointment system. However, most of the outpatient departments have multiple-providers and accommodate the patient's provider preferences to achieve patient satisfaction and continuity of care.

- Almost all the models assume that the patient is always available, while designing an appointment system and will accept any appointment slot that the clinic offers. In reality, patients might not be available at all times to visit a doctor.

  Therefore, while designing an appointment system (i.e., choice of appointment type and appointment rules), it is essential to incorporate the multi-phase nature of patient flow, patient preference and patient availability, to represent the actual conditions that exist in a healthcare system.

2.5.2 Opportunities to Improve the Accuracy of Appointment Schedules

The observations presented below are based on the review of literature on the environmental characteristics of the outpatient scheduling models (Section 2.3).

- Some papers assume that the entire demand (number of patient calls for appointments) to be known apriori and identify the appropriate appointment times for each patient.

Other papers assume that the demand to be random and construct the schedule, patient by patient, in a sequential fashion. The latter is most common in practice because the hospital does not exactly know the number of patients who will call for appointment on a given day.

- Most papers assume that the service times are independent and identically distributed. The service time distribution is mostly obtained using the historical data available.
- Most of the outpatient scheduling models assume that the patient no-show rate for all the patients to be equal to the average no-show rate of the clinic. However, the patient no-show value varies for each patient based on several characteristics, such as patient's age, appointment delay, weather conditions, etc. Therefore, patient no-show must be calculated for each patient, while evaluating different appointment rules.
- The patients are expected to arrive on-time in most of the scheduling models. However, patients may arrive late to their appointment and late arrivals increase the clinic's average patient waiting time and resource overtime.

It is important to consider different environmental factors to obtain realistic values for the performance measures used to evaluate the schedule. In other words, ignoring certain uncertainties might lead to improper evaluation of the schedule. Recent literature has addressed the uncertainty in demand and service time in their outpatient models. However, it is also important to include individual patient no-show value and patient lateness to improve the accuracy of evaluating the schedule.

2.5.3 Opportunities for Interactive Decision-Making Approach

The observations presented below are based on the review of literature on the outpatient scheduling model characteristics (Section 2.4).

- Almost all the outpatient scheduling models are developed as a single objective problem. The value of each performance measure is converted to an approximate dollar

value and the various performance measures are combined into a single objective (e.g., minimize total cost or maximize net profit). However, it would be difficult to estimate the equivalent dollar value for certain performance measures. For example, it is difficult to quantify the cost of unmet demand (i.e., patients who are not given an appointment in a given scheduling horizon) because the hospitals will not be able to quantify the loss of customer goodwill and the decrease in customer base. Similarly, it is difficult to quantify the dollar value of the waiting time of a patient because it depends on several factors, such as patient's occupation, age, and expectations.

- Some outpatient scheduling models have avoided the need to convert the various performance measures to dollar values by developing a multi-objective problem and presenting the set of non-dominated solutions. However, a large set non-dominated solutions increases the cognitive burden of the decision-makers (e.g., hospital administrators) and the decision-makers will not be able to decide the best compromise solution that suits their clinic characteristics.

- A multi-objective model that uses an interactive approach has not been addressed by any of the prior studies. The interactive approach is an iterative solution process in which the decision-maker continuously interacts to express his/her preferences for each solution until the best compromise solution is achieved.

Using multi-criteria interactive approach will eliminate the need to convert the value of a performance measure to a dollar value and will enable the model to obtain the most-preferred solution to the decision-maker without imposing much cognitive burden.

2.5.4 Opportunities to Study New Appointment Types

There is a growing trend to provide virtual appointments, where the patients are not required to visit the clinic for service. There are two type of virtual appointments - synchronous and asynchronous. Synchronous appointments require the presence of both the patient and the physician at the same time and a real-time communication between them. On

the other hand, asynchronous appointments involve the collection of medical data and transmitting them for off-line review by the physician. Hence, asynchronous appointments do not require both the patient and physician to be present at the same time.

The virtual appointments may increase patient's satisfaction because the patient can avoid unnecessary visits to the clinic and avoid long waiting times in the clinic's waiting room. However, to the best of the authors' knowledge, none of the studies in the literature include virtual appointments in the appointment system design. Therefore, a hybrid appointment system that incorporates virtual appointments could provide greater flexibility to the design of an appointment system.

Tables 2.7 and 2.8 gives a classification of the recent key publications in outpatient scheduling models that are relevant to this thesis. In addition, the contributions of this dissertation that addresses some of the gaps in the literature are also indicated.

Table 2.7: Classification of recent key publications based on appointment design and outpatient department characteristics

Author (Year)	**Data-driven appointment rules**	**Multi-phase patient flow**	**Provider preference**	**Patient availability**
Chakraborty et al. (2010)				✓
Gul et al. (2011)		✓		
Hahn-Goldberg et al. (2012)		✓		
Saremi et al. (2013)		✓	✓	
Yan et al. (2015)				✓
This Dissertation	✓	✓	✓	✓

Table 2.8: Classification of recent key publications based on environmental and model characteristics

Author (Year)	**Demand Uncertainty**	**Service time uncertainty**	**Individual patient no-show**
Chakraborty et al. (2010)	✓	✓	
Daggy et al. (2010)	✓	✓	✓
Lee and Yih (2010)			✓
Liu et al. (2010)	✓		✓
Gul et al. (2011)		✓	
Qu et al. (2011)			
Hahn-Goldberg et al. (2012)	✓	✓	
hline Patrick (2012)	✓		✓
Saremi et al. (2013)		✓	
Erdogan and Denton (2013)	✓	✓	
Peng et al. (2014)		✓	
El-sharo et al. (2015)			
Erdogan et al. 2015	✓	✓	
Yan et al. (2015)		✓	
This Dissertation	✓	✓	✓

2.6 Summary

In this chapter, a detailed review of the outpatient scheduling models in literature was presented in terms of the appointment design characteristics, outpatient department characteristics, environmental characteristics and model characteristics. Based on the review of literature, different opportunities for improvement were identified. They form the basis for the research in this dissertation.

Chapter 3

Analysis of the Impact of the Nature of Patient Flow and Patient Availability on the Schedule Outcomes

As discussed in Chapter 2, one of the major problems of the prior research work on designing an appointment system is that they assume simplified-clinic settings. For instance, most of the research work compute the total service time of a patient by adding his service times at each phase and then evaluate the schedule as a single-phase system. In reality, the patient moves through multiple stations or phases (i.e. check-in, visit nurse, visit physician, check-out) during their scheduled appointment visit. Hence, in a multi-phase system, patients may wait at each phase for a resource to become available. It is also possible for a resource at a phase to be idle, while waiting for the patient to complete service at an earlier phase. Therefore, the patient waiting time, resource idle time and overtime will be affected by the nature of the patient flow, even if the total service time of the patient in the system is constant. The appointment system that evaluates the schedule by combining all the service times, by ignoring the nature of patient flow, will be called the *simplified system*. The system that evaluates the schedule by considering the nature of patient flow, will be called the *flow-integrated system* throughout this chapter. Figure 3.1 shows a flow-integrated system with two phases and compares it to an equivalent simplified system assumed in the literature. In addition, most research work ignores the patient's availability and assume that all the patients, who call for an appointment, can be scheduled at any available time.

In reality, the process is complex and such appointment systems may no longer perform well when implemented in an actual system. Moreover, studies that focus on the nature of patient flow to improve the process ignore the design of the appointment system. In other words, they assume that the patients are already scheduled and therefore, do not evaluate the patient calls for appointment and appointment rules for scheduling a patient.

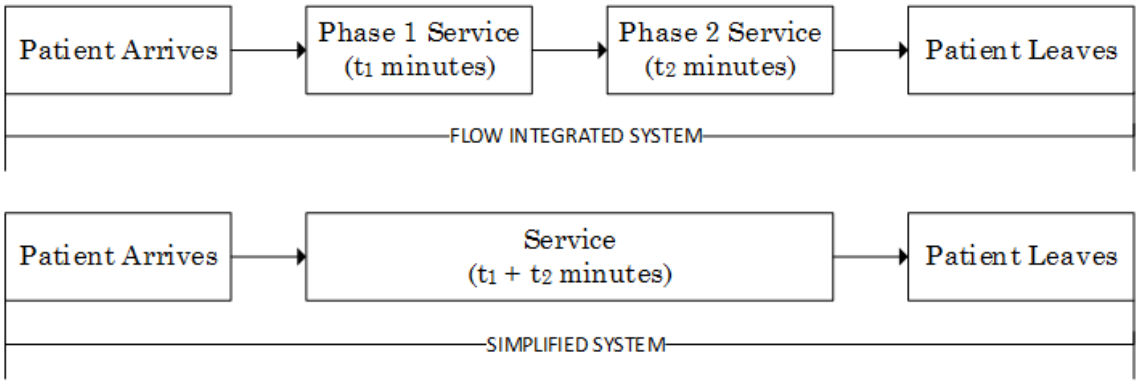


Figure 3.1: Comparison of flow-integrated and simplified system

Thus, the design of the appointment system and the nature of patient flow are rarely integrated in the literature. Hence, it is essential to identify their impact on the schedule outcomes (e.g., average waiting time, average idle time), so that the appointment rules proposed in the literature can be used in practice. In this chapter, we address this gap in the literature by developing an empirically based discrete event simulation model to evaluate the impact of patient flow on the schedule outcomes and determine whether the patient flow and patient availability must be integrated, while designing an appointment system.

3.1 Research Methodology

The research methodology involves two stages. In the first stage, a schedule is generated by assigning patients to a slot as and when the call occurs. In other words, the scheduling manager has to make an appointment decision for each patient before the end of his call. Therefore, the scheduling manager must decide without knowing the number of calls that he will encounter at a later time. The scheduling method, appointment rules, and the patient

availability determine the slot to which a patient can be scheduled and helps in decision making. Section 3.1.2 discusses the various appointment rules considered in this chapter.

In the second stage, the schedule generated in the first stage is evaluated for the two different systems, namely, the simplified system and the flow-integrated system. The schedule outcomes obtained from the two systems are statistically compared for each appointment rule to identify the impact of the nature of patient flow on the schedule outcomes of the appointment system.

3.1.1 Problem Description

A clinic must schedule a set of P patients over a period of D days based on an appointment rule. The operating hours of the clinic for each day in the scheduling horizon is constructed by appointment blocks and is termed as *slots*. Therefore, the number of slots for a given day is fixed based on the operating hours of the clinic. However, the duration of each slot varies depending upon the appointment rule used. For instance, IBFI rule has a constant appointment duration for all slots, whereas the OFFSET rule has varying appointment durations. The patient is scheduled to one of the slots in the scheduling horizon based on the patient characteristics, patient availability, and the appointment rules used by the clinic.

3.1.2 Appointment Rules and Model Assumptions

The various appointment rules studied in this chapter are given in Table 3.1. Based on the interaction with hospital administrators and the review of literature, several assumptions have been made:

- Some patients may not show-up for their appointment.
- The total number of patient calls for an appointment (demand) for a given day is greater than or equal to the total number of available slots (supply) on that day.
- The regular operating hours of the clinic is fixed.

- The clinic serves all the patient, who arrive for their scheduled appointments. Therefore, the clinic resources may work overtime in some situations.
- Walk-ins or any other unscheduled arrivals are not considered.
- Each phase has one server (resource).
- The total service time follows a known probability distribution.
- For multi-phase system (flow-integrated system), the service time of each phase is the product of workload parameter at that phase and total service time. The workload parameter for a phase is the average service time of that phase divided by the average service time of the hospital/clinic. For instance, if the average time spent at phase 1 is 10 minutes and the average service time of the clinic is 30 minutes, then the workload parameter for phase 1 is 10/30 = 1/3. The summation of workload parameter for all the phases must be equal to 1. The sum of service time at all the phases will be equal to the total service time.

Table 3.1: Appointment rules

Rule	**Description**
IBFI	Schedules one patient in each slot with a slot duration equal to the average service time
2ATBEG	Two patients are scheduled in the first slot and one patient is scheduled in each of the remaining slots
LVBEG	Patients with low service time variation are scheduled in the beginning and patients with high service time variation are scheduled at the end
HVBEG	Patients with high service time variation are scheduled in the beginning and patients with low service time variation are scheduled at the end
OFFSET	First $(k_1 - 1)$ patients have appointment duration less than the mean service time and the remaining patients have an appointment duration greater than the mean service time
DOME	First $(k_1 - 1)$ patients have an appointment duration less than the mean service time, patients $(k_1 + 1)$ to $(k_2 - 1)$ have an appointment duration greater than mean service time, and the rest of the patients have appointment duration less than mean service time

3.1.3 Sequence of events for each patient call

The sequence of events that take place for each patient call is as follows:

- Patient call for an appointment
- Patient communicates his availability with the scheduling manager (i.e., the potential days and slots at which he wishes to see his physician)
- Scheduling manager uses his prior knowledge and experience to classify the patient based on his service time variation, only if the clinic adopts an appointment rule that uses service time variation to schedule the patient.
- Scheduling manager scans each slot, for which the patient is available, to determine if the patient can be scheduled to one of these slots.
- The patient is scheduled only if a slot satisfies the restrictions imposed by the appointment rule. Otherwise, the patient cannot be scheduled in that slot. For example, most rules allow only one patient to be scheduled in a slot. Therefore, if the particular slot does not already have another patient scheduled, then the current patient can be scheduled in that slot.
- When the patient is scheduled, the scheduling manager confirms the appointment time and ends the call. If the patient cannot be scheduled to any of the slots, then the scheduling manager cannot schedule the patient in this scheduling horizon and asks the patient to call another time.

The same procedure is repeated for all the patients who call for an appointment and the schedule is constructed by adding patients to the schedule in accordance to the appointment rules. The final schedule includes the appointment information of each patient, namely, expected start time of each phase and the expected end time.

3.1.4 Sequence of events for schedule evaluation

The sequence of steps considered to evaluate the schedule for each day is as follows:

- If a patient does not show-up for his appointment, then the resources assigned to serve that patient will remain idle.
- For each phase, if the resource is available, then the arriving patient is served immediately upon arrival at that phase. If the resource is busy serving another patient who was scheduled earlier, then the arriving patient must wait until the resource becomes available. Therefore, the actual appointment start time (i.e., time at which the resource begins serving the patient) at a phase can be different from the expected appointment start time at that phase for that patient.
- After the resource becomes available, the patient is served for his entire service duration at that phase. The actual service completion time can be earlier, later or equal to the expected service completion time.
    - If the actual completion time is earlier than the expected completion time, then the resource is idle until the next appointment begins.
    - If the actual completion time is later than the expected completion time, then the patient in the next slot is expected to wait for the resource to be available.
    - If the actual completion time is the same as the expected completion time, then there is no idle time for the resource and no waiting time for the patient in the next slot.
- The patient exits the system after completing the service at all the phases.

The procedure is repeated for the next patient until all the patients in all the slots of the scheduling horizon are served. If the actual completion time of the last slot is later than the expected completion time, then it leads to overtime of the resources. The waiting time for

each patient, idle time of the resources at each slot and the overtime of the resources for each day are obtained.

3.1.5 Procedure for Schedule Construction (Algorithm 1)

The sequence of events discussed in Section 3.1.3 is followed for all the patients, who call for an appointment. The schedule is constructed by simulating the patient call-in process and adding the patient to a slot depending on the appointment rule under consideration. The step-by-step procedure associated with the schedule construction for a given scheduling horizon is referred to as *Algorithm 1* throughout this chapter. A high level flow-chart of Algorithm 1 is shown in Figure 3.2 and its detailed steps are described below:

Step 1: Select one appointment rule to schedule patients for the entire scheduling horizon.

Step 2: A patient calls for an appointment.

Step 3: Scan the first slot of the first day in the scheduling horizon.

Step 4: Check whether the slot can accommodate the patient. If the patient can be scheduled in the slot, then go to Step 5. Else, go to Step 8.

- For 2ATBEG rule, the first slot can accommodate two patients and all the other slots can accommodate one patient.
- For all the other appointment rules, each slot can accommodate only one patient.

Step 5: Check whether the patient can be scheduled in the slot for the selected appointment rule. If the patient can be scheduled in the selected slot, then go to Step 6. Else, go to Step 8.

- For LVBEG and HVBEG rule, if the patient is a low-variance patient and the slot is reserved for low-variance patients or if the patient is a high-variance patient and the slot is reserved for high-variance patients, then the patient can be scheduled in that slot.

- LVBEG rule schedules low-variance patients in the beginning and high-variance patients at the end. A fixed number of slots in the beginning are reserved for low-variance patients and the remaining slots are for high-variance patients.
- HVBEG rule schedules high-variance patients in the beginning and low-variance patients at the end. A fixed number of slots in the beginning are reserved for high-variance patients and the remaining slots are for low-variance patients.

- All other appointment rules schedule patient in any of the available slot.

Step 6: Check whether the patient is available to visit the clinic in the slot and day under consideration. If the patient is available, then go to Step 7. Else, go to Step 8.

Step 7: Schedule the patient in the slot and day under consideration, and go to Step 13.

Step 8: Check if the slot under consideration is the last slot of the day. If it is the last slot, then go to Step 10. Else, go to Step 9.

Step 9: Select the next slot and go to Step 4.

Step 10: Check if the day under consideration is the last day of the scheduling horizon. If it is the last day, then go to Step 12. Else, go to Step 11.

Step 11: Select the first slot of the next day, and go to Step 4.

Step 12: The patient cannot be scheduled in this scheduling horizon.

Step 13: Check if the patient under consideration is the last patient. If Yes, go to Step 15. Else go to Step 14.

Step 14: Wait for the next patient call, and go to Step 2.

Step 15: STOP. Schedule construction is complete.

A pseudo computer code for Algorithm 1 is given in the Appendix B.

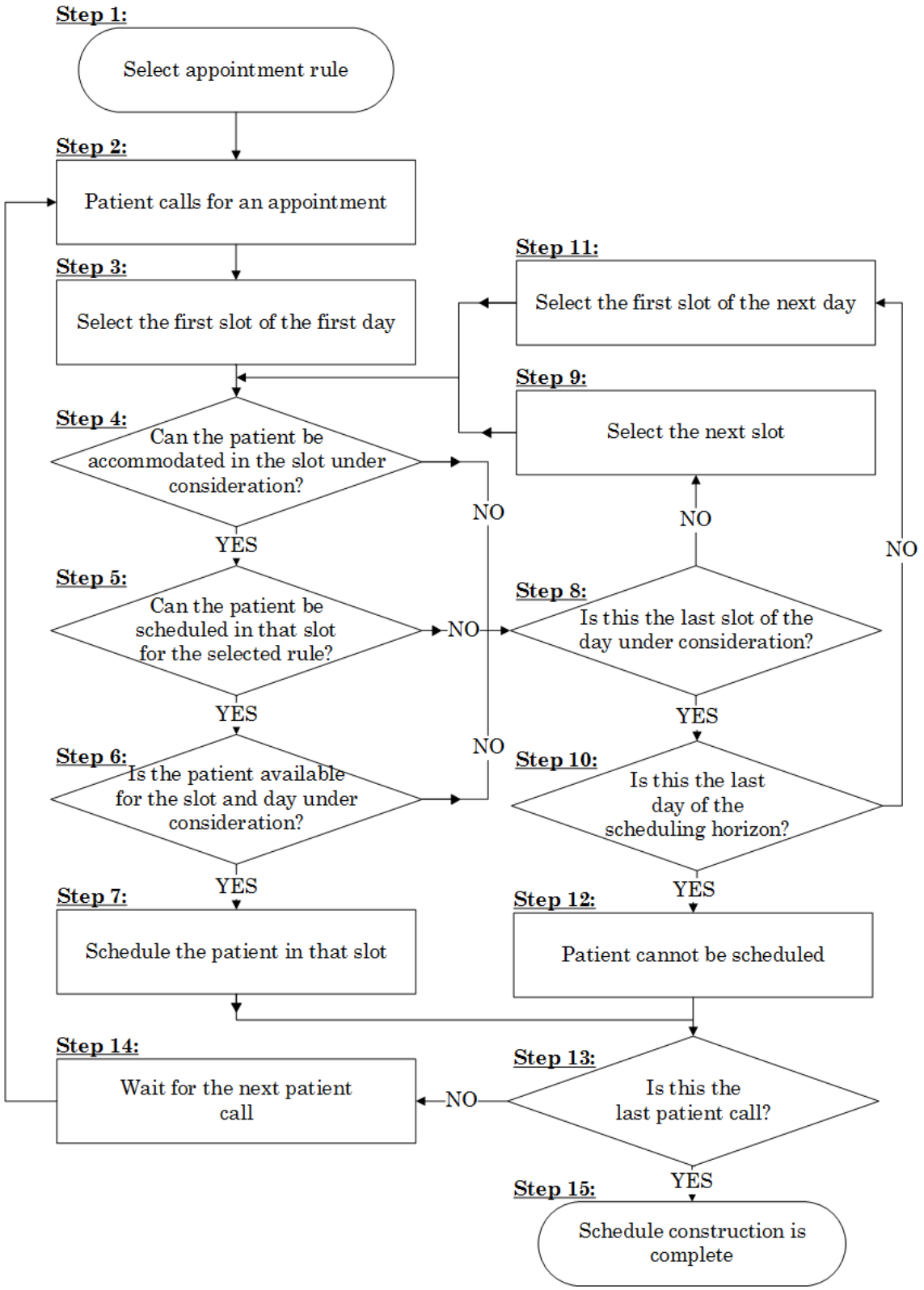


Figure 3.2: Flow-chart of Algorithm 1

3.1.6 Procedure for Schedule Evaluation (Algorithm 2)

The step-by-step procedure associated with the schedule evaluation is referred to as *Algorithm 2* throughout this chapter. Algorithm 2 simulates the patient arrivals to the clinic in accordance with the schedule generated by using Algorithm 1. The patient no-show value and their service times are sampled from known probability distributions. The distributions of patient no-show and service times can be obtained using the historical data. In order to incorporate the effect of uncertainty associated with patient no-show and service time, the schedule evaluation (Algorithm 2) is replicated several times. For each replication, the schedule generated by Algorithm 1 is evaluated by randomly sampling the no-show and service time value for each patient by Algorithm 2. Algorithm 2 is executed twice for each replication – one assuming a simplified system and another assuming a flow-integrated system.

Schedule Evaluation of a Simplified System

A high level flow-chart of Algorithm 2 for the simplified system is shown in Figure 3.3 and its detailed steps are described below:

Step 1: The fully constructed scheduled from Algorithm 1 is used to identify the slots to which each patient is scheduled.

Step 2: The evaluation for the first replication starts with the first slot of day 1.

Step 3: The actual service time and the no-show value (either 0 or 1) associated with all the patients scheduled in the slot under consideration is generated from known probability distributions. Even though, a patient may go through multiple phases (stations) in reality, it is approximated as a single phase in the simplified system for the schedule evaluation. In the flow-integrated system, each phase is modeled separately as illustrated in Figure 3.4.

Step 4: The actual start time of all the patients scheduled in the slot and day under consideration are determined. The service begins only when the resource is available, and the resource may be busy serving another patient. Therefore, the

actual start time can either be the appointment time or the time at which the resource is available. Further, the actual completion time of the patient in the selected slot is also determined. Section 3.1.7 discusses the events that affect the actual start time and completion time, and provides the mathematical expressions to determine them.

Step 5: Check if the selected slot is the last slot of the day under consideration. If it is the last slot, then go to Step 7. Else, go to Step 6.

Step 6: Select the next slot, and go to step 3.

Step 7: Check if the selected day is the last day of the scheduling horizon. If it is the last day, then go to Step 9. Else, go to Step 8.

Step 8: Select the first slot of the next day, and go to Step 3.

Step 9: Compute the schedule outcomes, namely patient waiting time, resource overtime and resource idle time, for the replication under consideration, using the actual start times and the actual completion times of all the patients scheduled in the scheduling horizon.

Step 10: Check if replication under consideration is the last replication. If YES, then proceed to Step 12. Else go to Step 11.

Step 11: Increment the replication value by one. Set slot and day values to one, and go to Step 3.

Step 12: Compute the average schedule outcomes by averaging across all replications. Section 3.1.7 discusses the schedule outcomes and provides the mathematical expressions to compute them.

Step 13: STOP. Schedule evaluation is complete.

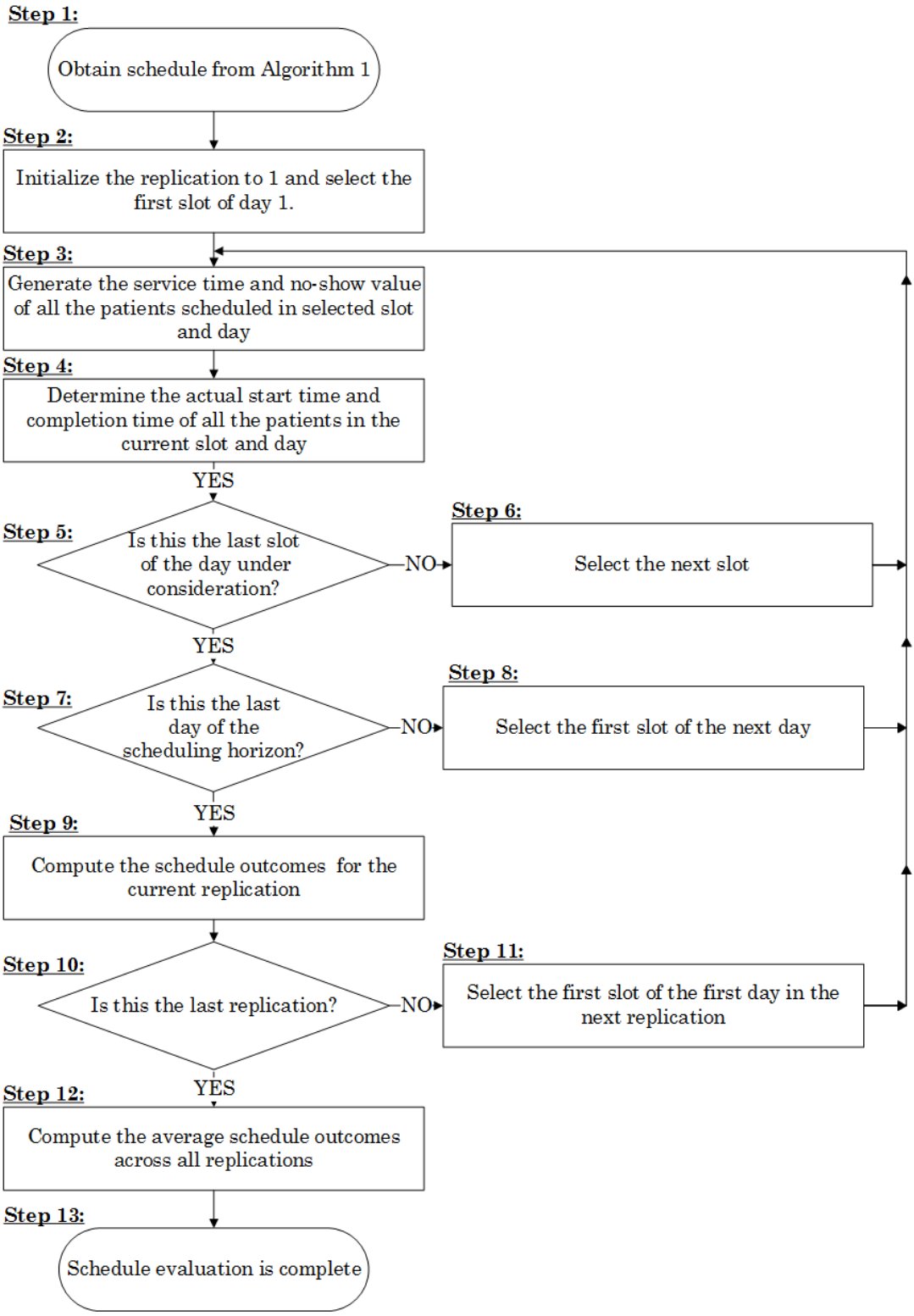


Figure 3.3: Flow-chart of Algorithm 2 for Simplified System

Schedule Evaluation of a Flow-Integrated System

The steps of Algorithm 2 for the flow-integrated system is similar to the simplified system but includes the evaluation of the schedule starting at each phase of a slot instead of each slot. A high-level flow-chart of Algorithm 2 for the flow-integrated system is shown in Figure 3.4. Here, the resource start-time and end-time are computed for each phase of a slot. Thus, the algorithm evaluates all the phases in a slot independently before proceeding to the next slot. However, the total service time of each patient is kept the same in both the simplified and flow-integrated systems.

A pseudo code for Algorithm 2 is given in Appendix C. Since the schedule for a given day may change after each patient call, the effectiveness of the schedule is evaluated by Algorithm 2 only after the schedule is fully constructed by Algorithm 1. The evaluation is first done by ignoring the patient flow in the clinic (simplified system) and then by considering the actual patient flow in the clinic (flow-integrated system). Thus, for each schedule, two different sets of schedule outcomes are obtained by Algorithm 2. The schedule outcomes obtained by ignoring the patient flow is compared statistically with those obtained using the actual patient flow.

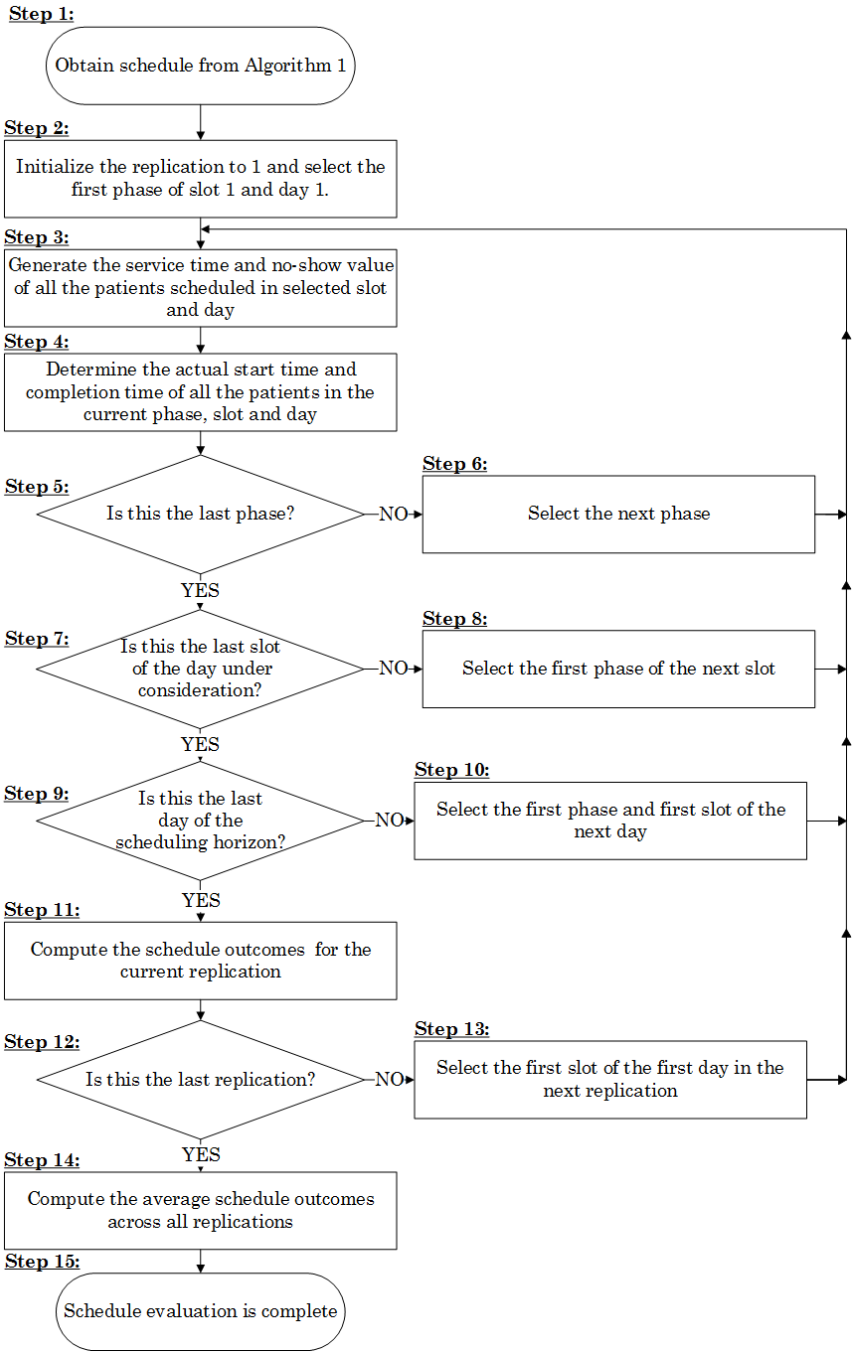


Figure 3.4: Flow-chart of Algorithm 2 for Flow-Integrated System

3.1.7 Schedule Outcomes

The notations used in this chapter are as follows. The notations are used to define the schedule outcomes. In addition, the basic steps of algorithms 1 and 2 using the notations are given in Appendix B and Appendix C, respectively.

Indices

$p, p', \hat{p}, \tilde{p}$	Patient index $\quad p, p', \hat{p}, \tilde{p} \in P$
s, s'	Slot index $\quad s, s' \in S$
d, d'	Day index $\quad d, d' \in D$
h	Phase index for $\quad h \in H$
i	Replication index $\quad i \in I$

Sets

P	Set of all scheduled patients
S	Set of all slots
D	Set of all days in the scheduling horizon
H	Set of all phases
I	Set of all replications
U	Set of all appointment rules

Parameters

$\alpha_{p,d,s}$	1 if patient p is available on day d in slot s; 0 otherwise
β_p	1 if service time variation of patient p is high; 0 otherwise
$\eta_{p,i}$	1 patient p shows-up for replication i; 0 otherwise
γ_h	Workload parameter for phase h
$ST_{p,h,i}$	Service time of patient p in phase h and replication i (in minutes)
S_1	Total number of slots for low-variance patients
S_2	Total number of slots for high-variance patients

Decision Variable

$A_{p,d,s}$	1 if patient p is scheduled on day d in slot s; 0 otherwise

Outputs

$EST_{h,s}$	Expected appointment start time in phase h and slot s
$AST_{p,h,i}$	Actual appointment start time of patient p in phase h and replication i
$ECT_{h,s}$	Expected appointment completion time in phase h and slot s
$ACT_{p,h,i}$	Actual appointment completion time of patient p in phase h and replication i
$WT_{p,i}$	Waiting time of patient p in replication i
$IT_{h,d,s,i}$	Idle time of resource in replication i, phase h, day d and slot s
$OT_{h,d,i}$	Overtime of resource in replication i, phase h and day d
$\overline{WT}$	Average waiting time of all patients across all replications
$\overline{IT}$	Average idle time of resources across all slots, phases, days and replication
$\overline{OT}$	Average overtime of resources across all phases, days and replication

The schedule outcomes evaluated in this chapter are as follows:

- average waiting time of the patients
- average overtime of the resources
- average idle time of the resources

In order to compute the schedule outcomes, it is essential to determine the expected start time, expected completion time, actual start time and actual completion time of each patient at each phase. The expected start time and the expected completion time for each phase can be calculated using the appointment duration of each phase. The appointment duration of phase h is the product of workload parameter for phase h (γ_h), and total appointment duration (τ). Consider a situation in which the appointment duration is 30 minutes, expected start time is 9:00 am and end time is 9:30 am. If the schedule is treated as a two-phase system and if $\gamma_1 = 1/3$ and $\gamma_2 = 2/3$ then, the appointment duration of the first phase is

$\frac{1}{3} \times 30 = 10$ minutes and the appointment duration of the second phase is $\frac{2}{3} \times 30 = 20$ minutes. The expected start time and end time for the first phase is 9:00 am and 9:10 am respectively, while the expected start time and end time for the second phase is 9:10 am and 9:30 am respectively.

For replication i, the actual start time of patient p in phase h ($AST_{p,h,i}$) is the maximum of four events: expected start time of phase h in slot s, actual completion time of patient p in phase $h-1$, actual completion time of another patient $\tilde{p}$ scheduled in slot $s-1$, actual completion time of an overbooked patient $\hat{p}$ scheduled in slot s and is given by Equation (3.1).

$$AST_{p,h,i} = \max\left(EST_{h,s}, ACT_{p,h-1,i}, ACT_{\tilde{p},h,i}, ACT_{\hat{p},h,i}\right) \tag{3.1}$$

For replication i, the actual completion time of patient p in phase h ($ACT_{p,h,i}$) is the sum of actual start time of patient p in phase h and the service time patient p in phase h as shown in Equation (3.2).

$$ACT_{p,h,i} = AST_{p,h,i} + ST_{p,h,i} * \eta_{p,i} \tag{3.2}$$

For replication i, the waiting time of patient p ($WT_{p,i}$) is the sum of the time between the actual service start time of patient p at phase h ($AST_{p,h,i}$) and the actual service completion time in phase $h-1$ ($ACT_{p,h-1,i}$) for all phases as shown in Equation (3.3).

$$WT_{p,i} = \sum_{h=1}^{H} \left(AST_{p,h,i} - ACT_{p,h-1,i}\right) \tag{3.3}$$

Note, that the actual completion time of phase 0 is equal to the appointment time of patient p (i.e., $ACT_{p,0,i}$=Appointment time). As shown in Equation (3.4), the average waiting time for the schedule is the summation of the waiting time for each patient across all replications divided by the total number of patients scheduled and the number of replications.

$$\overline{WT} = \frac{\sum_{i=1}^{I}\sum_{p=1}^{P} WT_{p,i}}{P*I} \tag{3.4}$$

A resource is idle only if the service completion of current slot (s) is earlier than the expected service completion time of that slot. Therefore, the idle time of the resource at phase h and slot s ($IT_{h,d,s,i}$), is the maximum of two numbers: 0 and the difference between expected service completion of phase h in slot s and actual service completion of phase $h-1$ in slot s. Equation (3.5) is the mathematical representation of the idle time for day d and replication i.

$$IT_{h,d,s,i} = \max\Big((ECT_{h,s} - ACT_{p,h-1,i}), 0\Big) \tag{3.5}$$

The average idle time (Equation 3.6) is the summation of the idle times at all phases, slots, replications and days divided by the total number of phases, replications, days and slots.

$$\overline{IT} = \frac{\sum_{i=1}^{I}\sum_{d=1}^{D}\sum_{h=1}^{H}\sum_{s=1}^{S} IT_{h,d,s,i}}{I*D*S} \tag{3.6}$$

A resource has to work overtime only if the actual completion time in the last slot (S) is later than the expected completion time of the last slot. Therefore, the overtime of resource at phase h on day d for replication i ($OT_{h,d,i}$) is given using Equation (3.7).

$$OT_{h,d,i} = \max\Big((ACT_{p,h,i} - ECT_{h,S}), 0\Big) \tag{3.7}$$

The average overtime (Equation 3.8) is the summation of overtimes at all replications, days, and phases divided by the total number of phases, replications and days.

$$\overline{OT} = \frac{\sum_{i=1}^{I}\sum_{d=1}^{D}\sum_{h=1}^{H} OT_{h,d,i}}{I*D} \tag{3.8}$$

The importance of integrating the nature of patient flow in the appointment system is evaluated using the schedule outcomes. The schedule outcomes obtained by considering

the actual nature of patient flow (i.e., flow-integrated system) is compared with the schedule outcomes obtained by ignoring the nature of patient flow (i.e., simplified system). If the difference between the schedule outcomes is not significant, then it can be concluded that simplified system is a good approximation to evaluate patient schedule. However, if the difference is significant, then the conclusions suggested in the literature, for that appointment rule, might not be true in reality for multi-phase systems.

3.2 Experimental Results

In this section, the model parameters are defined and the schedule is evaluated for both systems (simplified and flow-integrated) using different appointment rules. Further, the impact of patient no-show rate, service time distribution, and patient availability on the schedule outcomes for both systems are studied. The simulation models (Algorithm 1 and Algorithm 2) are coded and executed in C++ using a computer with an Intel Core i5 2.50 GHz processor with 8 GB of RAM.

3.2.1 Experimental Design

The experimental study is done by modeling a clinic with a scheduling horizon of 5 days and 8-hour work day. Six different appointment rules are studied to analyze the need for integrating patient flow in an appointment system. A summary of model parameters is given in Table 3.2 and are considered as the baseline setting in this chapter. The same model parameters are used as inputs for both the simplified and flow-integrated systems to compare their schedule outcomes. For example, the total service time of the patient is the same irrespective of the number of phases. The total service time for a patient is sampled from a known distribution. The service time for each phase is then computed by multiplying the workload parameter of a phase and total service time for that patient for the flow-integrated system.

Table 3.2: Summary of model parameters

Parameter	Value
Total number of replications (I)	500
Scheduling horizon (D)	5 days
Total number of slots per day (S)	16 slots
Number of phases for simplified system (H)	1
Simplified system: Work load parameter for phase 1 (γ_1)	1
Number of phases for flow-integrated system (H)	2
Flow integrated system: Work load parameter for phase 1 (γ_1)	1/3
Flow integrated system: Work load parameter for phase 2 (γ_2)	2/3
Average no-show rate of the system	0%
Patient availability	All slots
Service time distribution	Lognormal $\sim$ (30,5)
Service time range for low variance patients ($\beta_1 = 0$)	$25 \leq ST \leq 35$
Service time range for high variance patients ($\beta_1 = 1$)	$ST < 25$ and $ST > 35$
Total number of slots for low variance patients (S_1)	8 slots
Total number of slots for high variance patients (S_2)	8 slots
k value for OFFSET rule	8 slots
k_1 value for DOME rule	5 slots
k_2 value for DOME rule	9 slots

3.2.2 Analysis of Results

The baseline setting is used as the input for the simulation models (Algorithm 1 and Algorithm 2). Algorithm 1 is executed once to construct the appointment schedule for a given scheduling horizon by a particular appointment rule. Algorithm 2 is executed twice: initially for the simplified system (i.e., an actual multi-phase system that is approximated as a single-phase system) and then for the flow-integrated system (i.e., a multi-phase system without any approximations). The corresponding schedule outcome values (average patient waiting time, average resource overtime and idle time) are obtained. The changes in the schedule outcomes for the two systems are shown in Figure 3.5.

General Conclusions

Figures 3.5(a), 3.5(b) and 3.5(c) illustrate the impact of the nature of patient flow (simplified and flow-integrated system) on the average idle time, average overtime and average waiting time for different appointment rules. In addition to the average values, the standard deviations of the schedule outcomes across 500 replications are also shown in Figure 3.5. It is observed that the average resource idle time and its standard deviation for the flow-integrated system is higher compared to a simplified system, irrespective of the appointment rule. On the other hand, the average resource overtime and patient waiting time for a flow-integrated system is always lower compared to the simplified system irrespective of the appointment rule.

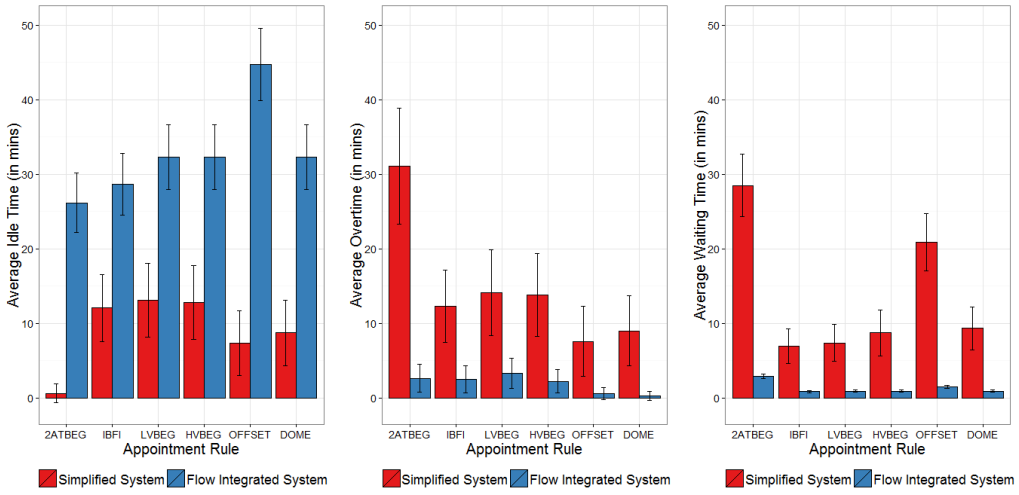


Figure 3.5: Schedule outcomes for simplified and flow-integrated system

Based on Figure 3.5, it is clear that the schedule outcomes for the two systems are different. However, their statistical significance is not obvious. Therefore, the Mann-Whitney test is performed at a significance level of 0.05 and it is found that the schedule outcomes of the two systems are significantly different (i.e., p value is less than 0.05). Hence, approximating a multi-phase system as a single phase for evaluating appointment rules is not valid.

Analysis of Resource Idle Time

Table 3.3 presents the average idle times for the various appointment rules and their rankings for both systems. The rankings of the appointment rules with respect to the idle time are different for the simplified system and the flow-integrated system. For example, the OFFSET rule is the second best rule to minimize the idle time for the simplified system, whereas it is the worst rule for the flow-integrated system. Moreover, the average resource idle times are significantly higher for the flow-integrated system.

Table 3.3: Average idle time (in mins) for various appointment rules and their associated rank for simplified and flow-integrated system

Appointment	**Simplified System**		**Flow-Integrated System**	
Rule	**Idle time**	**Rank**	**Idle time**	**Rank**
2ATBEG	0.62	1	26.15	1
IBFI	12.05	4	28.61	2
LVBEG	13.08	6	32.24	4
HVBEG	12.80	5	32.24	3
DOME	8.73	3	32.27	5
OFFSET	7.33	2	44.68	6

The IBFI rule resulted in the lowest increase with an average idle time value of the flow-integrated system being 2.4 times higher than the simplified system. The 2ATBEG rule resulted in the highest increase with an average idle time value of the flow-integrated system being 42 times higher than the simplified system. Therefore, the average idle time of the flow-integrated system is between 2.4 and 42 times higher than that of the simplified system. Some existing research work has concluded the 2ATBEG rule as one of the best rules to reduce the idle time of the clinic (Klassen and Rohleder, 1996; Ho and Lau, 1999; Kaandorp and Koole, 2007). When the nature of patient flow is integrated in the system, the increase in idle time value is highest for 2ATBEG rule, even though the ranking is still

the same.

Analysis of Resource Overtime

Table 3.4 presents the average overtimes for the various appointment rules and their rankings for both systems. The rankings of the appointment rules with respect to the overtime are slightly different for the simplified system and the flow-integrated system. Cayirli et al. (2006) identified the DOME rule to be inferior to the fixed-interval rules in terms of reducing the overtime. However, for the flow-integrated system, the DOME rule is the best when compared to the other appointment rules as it results in the least overtime. In addition, the average resource overtimes are significantly lower for the flow-integrated appointment system.

Table 3.4: Average overtime (in mins) for various appointment rules and their associated rank for simplified and flow-integrated system

Appointment	**Simplified System**		**Flow-Integrated System**	
Rule	**Overtime**	**Rank**	**Overtime**	**Rank**
2ATBEG	31.08	6	2.61	5
IBFI	12.30	3	2.49	4
LVBEG	14.10	5	3.29	6
HVBEG	13.81	4	2.23	3
DOME	8.98	2	0.30	1
OFFSET	7.57	1	0.58	2

The LVBEG rule resulted in the lowest decrease with an average overtime value of the flow-integrated system being 4.3 times lower than the simplified system, while the DOME rule resulted in the highest decrease with an average overtime value of the flow-integrated system being 30 times lower than the simplified system. In other words, the average overtime of the flow-integrated system is between 4.3 and 30 times lower than that of the

simplified system.

Analysis of Patient Waiting Time

Table 3.5 presents the average waiting times for the various appointment rules and their rankings for both systems. The ranking of the appointment rules does not change for the two systems. However, the average waiting times are significantly lower for the flow-integrated appointment system.

Table 3.5: Average waiting time (in mins) for various appointment rules and their associated rank for simplified and flow-integrated system

Appointment	**Simplified System**		**Flow-Integrated System**	
Rule	**Waiting Time**	**Rank**	**Waiting Time**	**Rank**
2ATBEG	28.47	6	2.86	6
IBFI	6.97	1	0.83	1
LVBEG	7.37	2	0.91	2
HVBEG	8.72	3	0.91	3
DOME	9.35	4	0.92	4
OFFSET	20.85	5	1.48	5

The LVBEG rule resulted in the lowest decrease with an average waiting time value of the flow-integrated system being 8 times lower than the simplified system, while the OFFSET rule resulted in the highest decrease with an average waiting time value of the flow-integrated system being 14 times lower than the simplified system. Therefore, the minimum decrease in the average waiting time for the flow-integrated system is 8 times the simplified system and the maximum decrease is 14 times the simplified system. Even though the OFFSET rule has a low waiting time, it is considered unacceptable by most existing research due to its high idle time (Klassen and Rohleder, 1996; Cayirli et al., 2006; Cayirli et al., 2012). In other words, this rule may be applicable only if the patient's waiting

time has a very high importance over the doctor's idle time. The results obtained for the OFFSET rule in this chapter is consistent with the existing research work. In addition, it is also found that the flow-integrated system magnifies the increase in idle time, as well as the decrease in overtime for the OFFSET rule.

Therefore, it is evident from the analysis that the recommendations suggested for a simplified-clinic system cannot be directly applied to the flow-integrated system, since the schedule outcomes of the two systems are significantly different. In other words, the appointment rules recommended in the literature cannot be applied to a real system that has multiple phases, where the nature of patient flow alters the schedule outcomes significantly.

3.2.3 Sensitivity Analysis

The baseline setting fixed certain parameters. In this section, some of the important parameters are varied to analyze their impacts on the schedule outcomes.

3.2.3.1 Impact of Service Time Distribution

The service time is sampled from a lognormal distribution in the baseline setting. However, in reality, it is possible for the clinics to experience a different service time distribution. The two other commonly observed service time distributions are exponential and uniform distribution (e.g., Muthuraman and Lawley, 2008; Peng et al., 2014; Yan et al., 2015). Hence, their impact on the schedule outcomes of the two systems is studied for different appointment rules. The mean of the exponential and uniform distribution is equal to the mean service time of lognormal distribution, and all the other parameters remain the same as the baseline setting.

Figures (3.6) - (3.8) indicate that the earlier conclusions, based on the initial analysis of the baseline setting, are consistent even if the service time distribution is exponential or lognormal. The idle time for flow-integrated system is significantly higher compared to the simplified system. The overtime and waiting time for flow-integrated system is significantly lower compared to the simplified system.

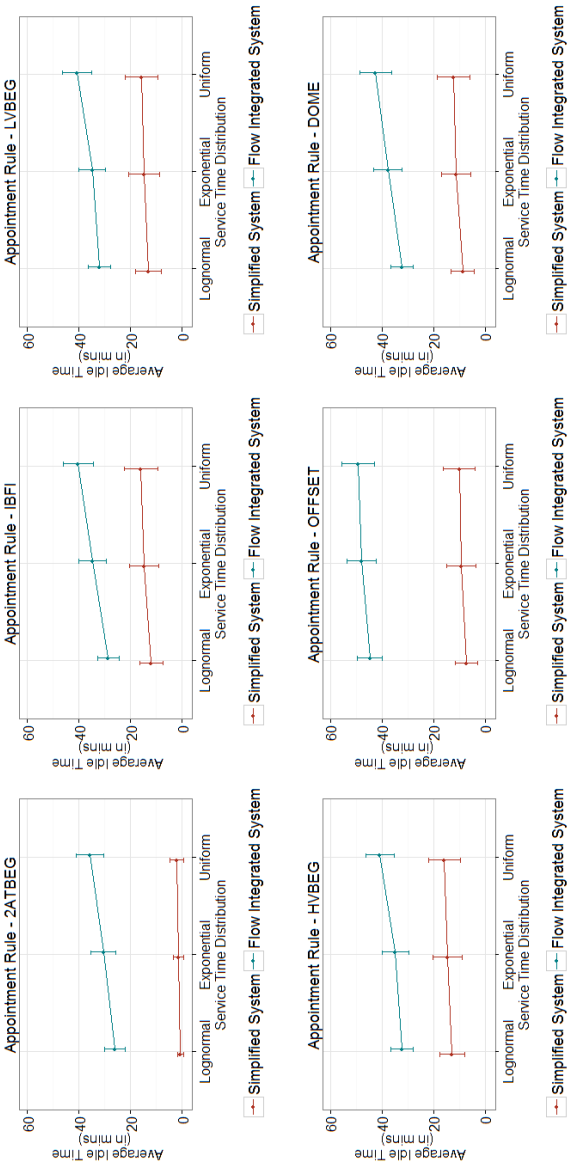


Figure 3.6: Impact of service time distribution on idle time of simplified and flow-integrated system

Irrespective of the appointment rule and type of system, it is evident from Figure 3.6 that the average idle time is low when the service time of patients is lognormally distributed and high when the service time is uniformly distributed. However, the standard deviation does not change significantly when the service time distribution changes.

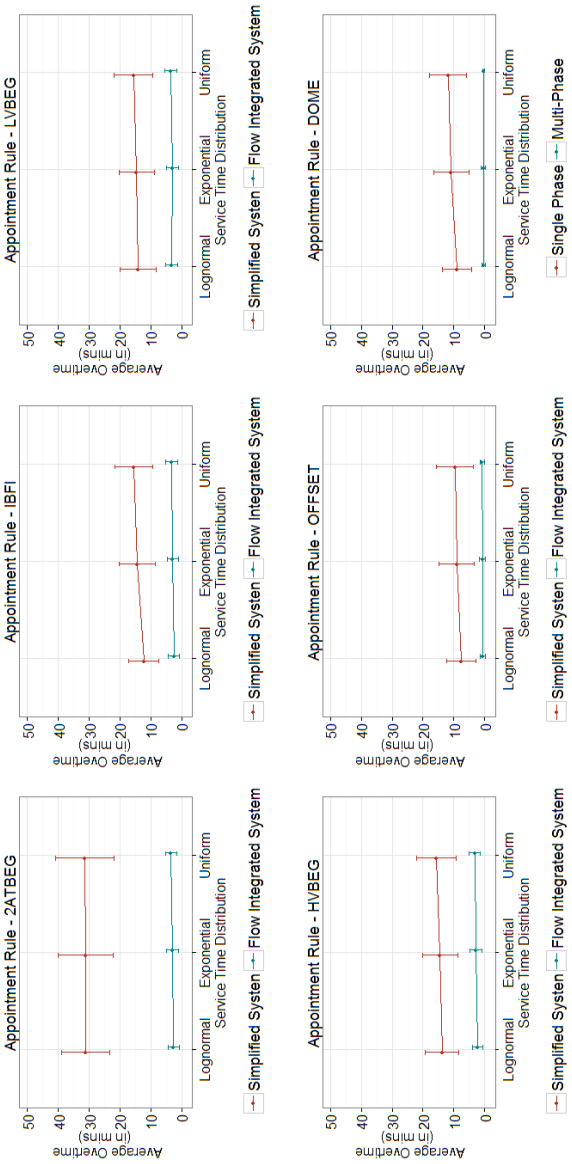


Figure 3.7: Impact of service time distribution on overtime of simplified and flow-integrated system

Irrespective of the appointment rule and type of system, it is clear from Figure 3.7 that the service time distribution does not have a major impact on the values of the average overtime and its standard deviation.

Similar to overtime, Figure 3.8 indicates that the service time distribution does not have a major impact on the values of the average waiting time and its standard deviation. Therefore,

the service time distribution mainly impacts the idle time of the resources, irrespective of the appointment rule or type of system.

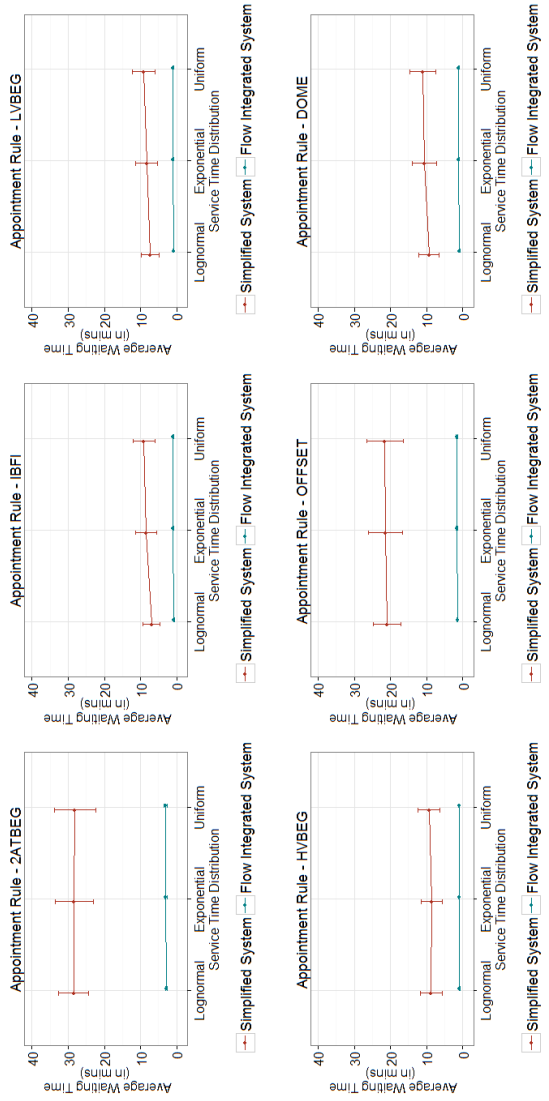


Figure 3.8: Impact of service time distribution on waiting time of simplified and flow-integrated system

3.2.3.2 Impact of No-show Rate

In the baseline setting, the average no-show rate is 0% (i.e., all the patients show-up for their appointment). According to one study, the national average for no-show rate was 5.5% in 2000 (Izard, 2005). Some outpatient departments experience high no-show rates of up to

34% (Geraghty et al. 2007, Dreiher et al. 2008). Therefore, the impact of the no-show rate on the two systems is studied in this section. The no-show rate is varied from 0% to 40%, in increments of 10%. If the average no-show rate is 10%, then on an average, 10% of the patients do not show-up for their scheduled appointment.

Impact of No-show Rate on Average Resource Idle Time

Figure 3.9 illustrates the impact of the average no-show rate on the average idle times for the simplified system and the flow-integrated system. Irrespective of the type of system and the appointment rules, the average idle time increases as the average no-show rate increases. This is obvious because the resource remains idle, when the patient does not show-up and none of the appointment rules adopt any overbooking policy to compensate for no-shows. It is interesting to note that the increase in average idle time is linear for the both systems.

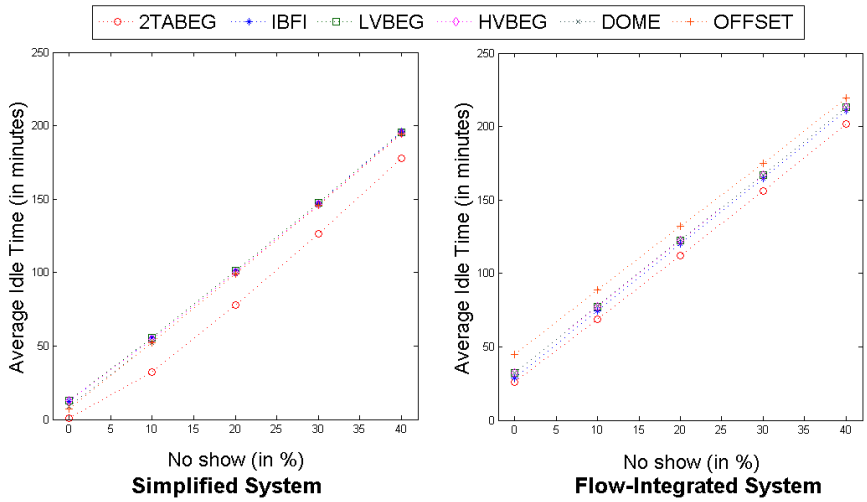


Figure 3.9: Impact of no-show rate on the resource idle time of simplified and flow-integrated system

Table 3.6 gives the average resource idle times for both systems under different appointment rules and no-show rates. It is clear that the idle times are always higher for the flow-integrated system compared to the simplified system for all no-show rates. Moreover, the difference between the idle time values decreases as the no-show rate increase. For the 2ATBEG rule, the average idle time for the flow-integrated system is 42 times higher

than the simplified system, when the no-show rate is 0% and only 1.1 times higher than the simplified system, when the no-show rate is 40%. However, the differences between the average values of idle time between the two systems are still significant.

Table 3.6: Impact of patient no-show rate on resource idle time (in mins)

Appointment	**Simplified System**					**Flow-Integrated System**				
Rule	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**
2ATBEG	0.62	32.30	77.94	126.23	177.52	26.15	68.81	112.38	155.83	201.62
IBFI	12.05	55.47	101.23	147.51	195.81	28.61	74.49	119.95	164.73	211.00
LVBEG	13.08	55.58	101.41	147.34	195.74	32.24	77.12	122.63	166.92	212.93
HVBEG	12.80	54.39	100.42	146.53	195.00	32.24	77.10	122.62	166.91	212.91
DOME	8.73	52.44	98.87	145.63	194.34	32.27	77.74	122.82	167.20	213.06
OFFSET	7.33	52.44	99.02	145.63	194.34	44.68	88.53	132.03	175.00	219.47

Impact of No-show Rate on Average Resource Overtime

Figure 3.10 illustrates the impact of the average no-show rate on the average overtimes for the simplified system and the flow-integrated system. Irrespective of the type of system and the appointment rules, the average overtime decreases as the average no-show rate increases. In other words, as the probability of a patient not showing up increases, the possibility of a resource working beyond the clinic operating hours will decrease.

Table 3.7 gives the average resource overtimes for both systems under different appointment rules and no-show rates. It is clear that the average overtime for the flow-integrated system is always lower compared to the simplified system for all no-show rates. In addition, the difference in overtime values for the flow-integrated system and simplified system decreases as the average no-show rate increases. For the 2ATBEG rule, the average overtime for the flow-integrated system is 12 times lower than the simplified system, when the no-show rate is 0% and only 1.4 times lower than the simplified system, when the no-show rate is 40%. However, the differences between the average overtime values of the two systems are still significant.

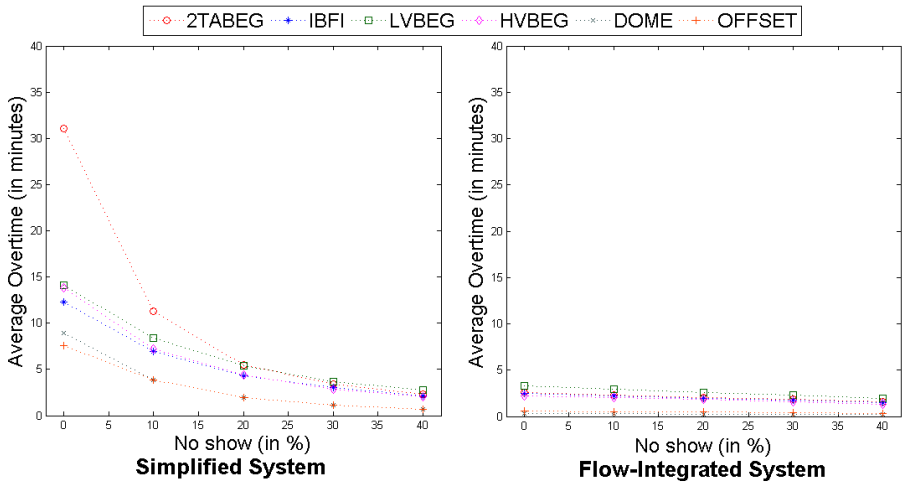


Figure 3.10: Impact of no-show rate on the average resource overtime of the simplified and flow-integrated system

Table 3.7: Impact of patient no-show rate on resource overtime (in mins)

Appointment	**Simplified System**					**Flow-Integrated System**				
Rule	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**
2ATBEG	31.08	11.26	5.52	3.42	2.34	2.61	2.30	2.06	1.85	1.58
IBFI	12.30	6.91	4.35	3.03	2.18	2.49	2.23	1.97	1.75	1.50
LVBEG	14.10	8.37	5.35	3.71	2.78	3.29	2.94	2.59	2.28	1.98
HVBEG	13.81	7.18	4.36	2.90	2.04	2.23	2.02	1.82	1.59	1.34
DOME	8.98	3.87	1.95	1.15	0.71	0.30	0.27	0.23	0.22	0.19
OFFSET	7.57	3.87	1.91	1.15	0.71	0.58	0.52	0.45	0.42	0.36

Impact of No-show Rate on Average Patient Waiting Time

Figure 3.11 illustrates the impact of the average no-show rate on the average patient waiting times for the simplified system and the flow-integrated system. The impact of no-show rate on the average waiting time is similar to the impact of the no-show rate on the average overtime. Irrespective of the type of system and appointment rules, the average waiting time decreases as the average no-show rate increases.

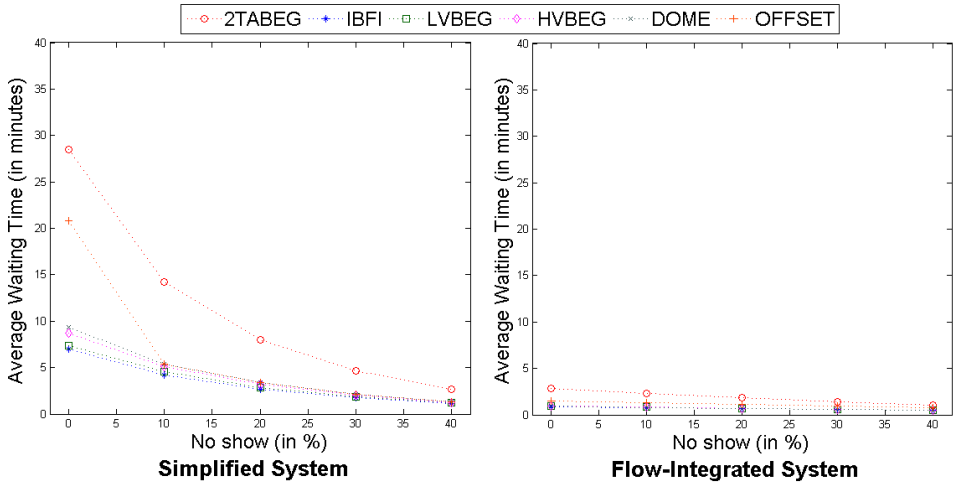


Figure 3.11: Impact of no-show rate on average patient waiting time of the simplified and flow-integrated system

Table 3.8 gives the average patient waiting times for both systems under different appointment rules and no-show rates. It is clear that the average waiting time for the flow-integrated system is always lower compared to the simplified system for all no-show rates. Similar to the resource idle time and overtime, the difference between the waiting times for the flow-integrated system and simplified system decreases as the average no-show rate increases.

Table 3.8: Impact of patient no-show rate on patient waiting time (in mins)

Appointment	**Simplified System**					**Flow-Integrated System**				
Rule	**0%**	**10%**	**20%**	**30%**	**40%**	**0%**	**10%**	**20%**	**30%**	**40%**
2ATBEG	28.47	14.25	7.97	4.61	2.70	2.86	2.30	1.84	1.44	1.08
IBFI	6.97	4.21	2.67	1.72	1.11	0.83	0.73	0.65	0.56	0.48
LVBEG	7.37	4.56	2.88	1.89	1.21	0.91	0.81	0.72	0.62	0.53
HVBEG	8.72	5.12	3.19	2.01	1.26	0.91	0.82	0.72	0.63	0.53
DOME	9.35	5.39	3.34	2.13	1.35	0.92	0.81	0.71	0.62	0.52
OFFSET	20.85	5.39	3.37	2.13	1.35	1.48	1.30	1.13	0.97	0.81

For the 2ATBEG rule, the average waiting time for the flow-integrated system is 10 times

lower than the simplified system, when the no-show rate is 0% and only 2.5 times lower than the simplified system, when the no-show rate is 40%. However, the differences between the average waiting time values of the two systems are still significant.

It is evident from the analysis that the gap between the schedule outcomes for the two systems decreases as the no-show rate increases. However, the differences in their average values are still significant. Therefore, to obtain realistic schedule outcomes, the nature of patient flow must be integrated in the design of an appointment system, even if the no-show rate is 40%.

3.2.3.3 Impact of Patient Availability

The clinic is assumed to operate for 8 hours (excluding the one hour lunch break) and the operating hours are categorized into four sessions, namely, morning (8AM–10AM), forenoon (10AM–12PM), afternoon (1PM–3PM) and evening (3PM–5PM). In the baseline setting, the patient accepts any slot that is offered to him. In other words, the patient is available to visit the clinic during all the four sessions. In this section, four different availability settings are studied to understand the impact of patient availability on the two systems for different appointment rules. In the first three availability settings (A1-A3), the patient's availability is restricted to two of the four sessions. In availability setting A4, the patient's availability is restricted to just one of the four sessions, and hence, is the most restrictive setting among all the availability settings.

In availability setting A1, a calling patient is available either during 8AM–10AM & 1PM–3PM or during 10AM–12PM & 3PM–5PM. In availability setting A2, a patient is available either during 8AM–10AM & 3PM–5PM or during 10AM–12PM & 1PM–3PM. In availability setting A3, each patient is available either during 8AM–10AM & 10AM–12PM or during 1PM–3PM & 3PM–5PM. In availability setting A4, every patient is available only during one of the sessions, namely 8AM–10AM, 10AM–12PM, 1PM–3PM, or 3PM–5PM.

Tables (3.9) through (3.11) present the schedule outcomes under different availability settings and appointment rules. In the Baseline setting, the patient is available at any time

during the day. It is clear that the schedule outcomes for all the appointment rules, except for the LVBEG and HVBEG rule, are not affected by the availability settings, both for the simplified system and the flow-integrated system. For the LVBEG and HVBEG rule, the resource idle time increases and the resource overtime decreases for availability settings A3 and A4. In all the other cases, the values of the resource idle times and overtimes, as well as the patient waiting times, are nearly the same as in the baseline setting, where the patient is available for all 4 sessions.

In addition, for each appointment rule, the differences between the schedule outcome values of the two systems, under all the availability settings, are almost equal to differences between the schedule outcome values of the two systems under the Baseline setting. This is because the change in value of the schedule outcome for both systems under any availability setting (A1-A4) is almost equal to the change in value of the schedule outcome for both systems under the Baseline setting.

Table 3.9: Impact of patient availability settings on resource idle time

Appointment	**Simplified System**					**Flow-Integrated System**				
Rule	**Baseline**	**A1**	**A2**	**A3**	**A4**	**Baseline**	**A1**	**A2**	**A3**	**A4**
2ATBEG	0.62	0.61	0.66	0.58	0.54	26.15	26.09	26.25	26.38	26.67
IBFI	12.05	12.14	11.92	11.96	11.73	28.61	28.90	28.63	28.86	29.06
LVBEG	13.08	13.24	13.13	15.96	15.87	32.24	32.31	32.35	37.47	37.45
HVBEG	12.80	12.85	12.80	34.05	36.90	32.24	32.28	32.24	54.99	60.29
DOME	8.73	8.78	8.60	8.75	8.52	32.27	32.63	32.18	32.42	32.53
OFFSET	7.33	7.48	7.32	7.53	7.36	44.68	45.01	44.66	45.13	45.23

Table 3.10: Impact of patient availability settings on resource overtime

Appointment	Simplified System					Flow-Integrated System				
Rule	Baseline	A1	A2	A3	A4	Baseline	A1	A2	A3	A4
2ATBEG	31.08	31.14	31.00	31.02	31.24	2.61	2.41	2.60	2.69	2.80
IBFI	12.30	12.32	12.29	12.27	12.14	2.49	2.23	2.53	2.55	2.66
LVBEG	14.10	14.23	14.15	11.18	11.21	3.29	3.41	3.29	2.85	2.74
HVBEG	13.81	13.84	13.81	10.96	7.74	2.23	2.07	2.19	1.84	1.74
DOME	8.98	8.97	8.97	9.07	8.93	0.30	0.23	0.34	0.29	0.32
OFFSET	7.57	7.67	7.68	7.85	7.76	0.58	0.48	0.62	0.57	0.64

Table 3.11: Impact of patient availability settings on patient waiting time

Appointment	Simplified System					Flow-Integrated System				
Rule	Baseline	A1	A2	A3	A4	Baseline	A1	A2	A3	A4
2ATBEG	28.47	28.42	28.53	28.63	28.79	2.86	2.84	2.85	2.86	2.87
IBFI	6.97	7.15	7.01	7.08	7.08	0.83	0.83	0.83	0.83	0.84
LVBEG	7.37	7.38	7.36	7.16	7.37	0.91	0.91	0.91	0.91	0.90
HVBEG	8.72	8.73	8.74	7.52	5.39	0.91	0.92	0.91	0.92	0.91
DOME	9.35	9.46	9.36	9.48	9.48	0.92	0.93	0.92	0.93	0.93
OFFSET	20.85	21.00	20.97	21.13	21.19	1.48	1.48	1.48	1.50	1.49

Table 3.12 gives the number of unscheduled slots (i.e., slot in which no patients are scheduled) obtained using Algorithm 1, under different availability settings and appointment rules. Baseline and availability settings A1 and A2 do not have any unscheduled slots for any of the appointment rules. Further, availability settings A3 and A4 do not have any unscheduled slots for 2ATBEG, IBFI, DOME and OFFSET appointment rules. In other words, all slots are filled with patients, increasing resource utilization and patient waiting times. However, for settings A3 and A4, LVBEG rule has one unscheduled slot. Similarly, the HVBEG rule has four unscheduled slots for setting A3 and five unscheduled slots for setting A4. This is mainly because of the additional restrictions imposed by the LVBEG

and HVBEG rules.

Table 3.12: Impact of patient availability settings on the schedule

Appointment	**Number of Unscheduled Slots**				
Rule	**Baseline**	**A1**	**A2**	**A3**	**A4**
2ATBEG	0	0	0	0	0
IBFI	0	0	0	0	0
LVBEG	0	0	0	1	1
HVBEG	0	0	0	4	5
DOME	0	0	0	0	0
OFFSET	0	0	0	0	0

Recall that the LVBEG rule schedules the low-variance patients in the first two sessions and high-variance patients in the last two sessions, while the HVBEG rule schedules the high-variance patients in the first two sessions and low-variance patients in the last two sessions. Moreover, in setting A3, a patient is available either in the first two sessions or in the last two sessions and in setting A4, the patient is available only during one of the four sessions. Therefore, if a high variance patient is available in the first two sessions, then that patient cannot be scheduled under the LVBEG rule. Similarly, if a low variance patient is available only in the first sessions, then that patient cannot be scheduled under the HVBEG rule. Therefore, the patient availability and the slot to which the patient can be scheduled do not match, which results in unscheduled slots. As a result, the schedule outcomes changed for both LVBEG and HVBEG rules under both systems using Algorithm 2.

It can be observed from Table 3.9 and Table 3.12 that the average resource idle time is nearly the same when there are no unscheduled slots and the average resource idle time increases in the presence of unscheduled slots. Similarly, it can be observed from Table 3.10 and Table 3.12 that the average resource overtime is nearly the same when there are no unscheduled slots and the average overtime decreases in the presence of unscheduled slots.

Unlike the other appointment rules, LVBEG and HVBEG rules restrict the type of patients who can be scheduled in a slot. In addition, the patient availability also restricts the slot to which the patient can be scheduled. Therefore, it is possible to have unscheduled slots when there are too many restrictions on scheduling a patient to a time slot.

3.3 Conclusions

There are different appointment rules that have been proposed in the literature to improve the schedule outcomes, such as patient waiting time, resource idle time. Almost all the existing research work ignore the multi-phase nature of patient flow and consider a simplified single-phase clinic system (actual multi-phase system approximated to a single-phase system) to evaluate the schedule. Further, most of the research work do not consider the impact of patient availability on the schedule outcomes. However, most clinic's patient flow involves multiple phases and the patients may be available only during certain times to visit the hospital. In this chapter, the impact of the nature of patient flow and patient availability on the schedule outcomes are evaluated for six different appointment rules.

First, the schedule is constructed by assigning patients to slots as and when they call for an appointment. The schedule generation depends on the appointment rule and the patient availability. Once the schedule is fully constructed, the schedule outcomes (average patient waiting time, average resource idle time and average resource overtime) are calculated by considering two different system settings, namely, simplified system and flow-integrated system. The simplified system ignores the multi-phase nature of patient flow while evaluating the schedule outcomes and it is the most commonly modeled system in the literature. On the other hand, the flow-integrated system considers the nature of patient flow to compute the schedule outcomes. Further, the same model parameters (e.g., patient availability, total service time) are used as inputs for both the system settings to compare their schedule outcomes. Algorithm 1 and Algorithm 2 proposed in this chapter are used in the schedule construction and outcome evaluation simulation models, respectively. The proposed algorithms are generic and can be adapted to different parameter settings.

The initial analysis (baseline setting) indicated that the average idle time, average overtime and the average waiting time are significantly different for the two systems. The average idle time is low for the simplified system and comparatively high for the flow-integrated system. On the other hand, the average overtime and waiting time are high for the simplified system and are comparatively low for the flow-integrated system. Therefore, the recommendations on the appointment rules to improve the schedule outcomes based on simplified systems are not valid for multi-phase systems. They have to be studied as flow-integrated systems to get valid results.

The model parameters are then varied to study the impact of service time distribution, patient no-show, and patient availability on the schedule outcomes of the two systems. Compared to the baseline setting, the average idle time increases for both systems when the service time distribution is uniform or exponential. However, the service time distribution did not have any significant impact on the average overtime and waiting time. In addition, the schedule outcomes of the two systems were significantly different even for different service time distributions.

The increase in no-show rate resulted in a linear increase in average idle time and an exponential decrease in average overtime and waiting time for the simplified system. On the other hand, the increase in no-show rate resulted in a linear increase in average idle time and a linear decrease in average overtime and waiting time for flow-integrated system. Moreover, the increase in no-show rate decreases the gap between the schedule outcomes for the two systems. The analysis indicated that the schedule outcomes of the two systems are significantly different even when the no-show rate was 40%. However, the difference between the average values of the two systems may become insignificant when the average no-show rate exceeds a certain value. In such situations, the schedule outcomes based on the simplified system may be a good approximation to the flow-integrated system.

In addition to the baseline setting, the impact of schedule outcomes on the two systems for four different patient availability setting was also studied. In both systems, availability setting A3 (patient is available either during 8AM-10AM & 10AM–12PM or during 1PM–

3PM & 3PM–5PM) and availability setting A4 (patient is available only during one of the four sessions, namely, 8AM-10AM, 10AM–12PM, 1PM–3PM, 3PM–5PM) lead to unscheduled slots for LVBEG and HVBEG appointment rules. As a result, the schedule outcomes for both systems are impacted resulting in unscheduled time slots. However, the impacts were minimal for both systems. For all the other appointment rules, the availability settings had no impact on the unscheduled slots.

It is evident from the analysis that the schedule outcomes are significantly different for the simplified and flow-integrated system. Therefore, the recommendations on the appointment rules using a simplified clinic system cannot be directly applied to a multi-phase clinic. Hence, the appointment rules that are known to improve the schedule outcomes in the literature may not necessarily perform well under realistic conditions (multi-phase patient flow system). This emphasizes the need for integrating the nature of patient flow and patient availability when designing an appointment system.

The analysis indicated that patient no-show rate leads to an increase in average idle time for all the appointment rules. Further, it is observed that incorporating patient availability may lead to unscheduled slots for certain appointment rules, thereby increasing the idle time of resources. Therefore, it is necessary to develop an appointment system that improves the schedule outcomes by taking into account the patient no-shows and patient availability. This is done in Chapter 4.

Chapter 4

Designing Multi-Phase Multi-Provider Hybrid Appointment Systems under Uncertainty

Most recent research focus on designing a hybrid appointment system (a combination of two or more appointment types) because of its potential to achieve the advantages of more than one appointment type. The primary objectives of the appointment system are to achieve patient satisfaction and effective resource utilization. If a clinic adopts a hybrid appointment system by combining open access (OA) and pre-booking (PB), then the OA can reduce the appointment delay leading to higher patient satisfaction, and the PB can provide steady patient flow leading to better resource utilization. However, to achieve the best outcomes, it is essential to determine the number of slots reserved for each appointment type (i.e., OA and PB). Overestimating the number of slots reserved for OA appointments affects the number of PB patients scheduled and the resource utilization, while underestimating the number of slots reserved for OA appointments affects the number of OA patients scheduled. Further, the uncertainty associated with the patient's service time, patient no-show rate, demand for both PB and OA patients' complicates the task of estimating the number of slots for each appointment type.

Based on the review of literature presented in Chapter 2, most of the existing research consider a single phase single provider system (hospital or clinic). Further, patients are scheduled without considering their availability. However, in practice, patients have to

go through multiple phases (pre-screening, treatment, and checkout) and the patients are available only during certain time slots. The analysis in Chapter 3 indicates that ignoring the actual nature of patient flow will alter the values of average waiting time, average resource overtime and idle time. Further, the presence of patient availability may alter the schedule and may lead to unscheduled slots. Therefore, an appointment system that does not integrate the actual nature of patient flow does not represent the real situation and therefore, cannot be applied to a real system that has multiple phases. In addition, patient availability plays a critical role in outpatient scheduling and the existing schedule that was obtained ignoring patient availability may no longer be feasible. Further, most of the research ignores the workload associated with a resource, and the uncertainty associated with patient calls for appointment, patient service time, patient no-show, patient availability and patient's provider preference, while scheduling a patient to a particular slot.

To address these gaps in the literature, a mathematical model for a hybrid appointment system (combination of OA and PB) that provides service to patients in multiple phases is proposed in this chapter. The model is developed with the objective of minimizing the total cost (i.e., sum of the opportunity cost due to an unscheduled patient, patient waiting time cost, resource idle time cost and resource overload cost). Further, the model developed in this chapter is generic and can be applied to any outpatient clinic by modifying the system parameters such as working hours of the clinic, and arrival rate of the patients. The proposed model aids the hospital administration by providing information on the number of OA and PB slots to be reserved for a given day, their position in the schedule, and the position of slots to be double-booked by taking into account the nature of patient flow, patient availability, patient's provider preference, and uncertainties associated with demand, patient service time and patient no-show rate.

4.1 Methodology

4.1.1 Problem Description

In this chapter, we consider an outpatient clinic that adopts a hybrid appointment system (combination of OA and PB). Further, it is assumed that the clinic provides service to patients in two phases as shown in Figure 4.1. Therefore, the patient flow involves a sequential movement of an arriving patient through two phases before exiting the system. At each phase, there may be multiple resources. Further, each resource has its own schedule and therefore, one resource does not serve the patients scheduled for another resource. The schedule for each resource is made of constant time periods called slots and the patients calling for an appointment may be scheduled to one of the slots. However, if a patient is scheduled, then he must be scheduled to a resource at each phase of the multi-phase system.

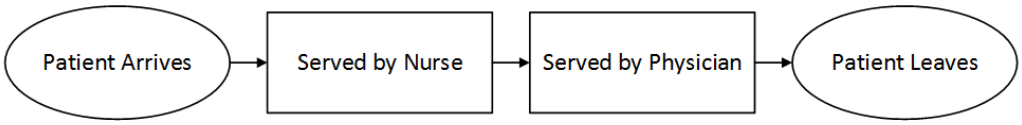


Figure 4.1: Patient flow of the outpatient clinic under study

The slots are classified as PB slots and OA slots based on the patient type that they accommodate. A patient who is not scheduled to any of the slots is considered as an unscheduled patient throughout this chapter. Patients call for appointments for a given day (i.e., a specific date). Therefore, a PB patient who calls for an appointment for a given day can occur weeks or months in advance, while an OA patient who calls for an appointment for a given day can occur only on that day. For example, if the day under consideration has 10 PB requests and 5 OA requests, then the patient calls for the 10 PB requests could have happened any day prior to that day under consideration. However, patient calls for the 5 OA requests would happen only on the day under consideration. Also, the patient calling for an appointment may have a provider preference and may only be available during certain slots. The day on which the patient called for an appointment affects the patient's no-show rate as

it is dependent on the appointment delay. A patient with longer appointment delay has a higher no-show rate. Moreover, patient service time is not constant as each patient might require different procedures.

4.1.2 Scenario for a Typical Day

For a given day, the clinic has certain number of OA appointment requests and PB appointment requests. Further, each request has several attributes (patient service time, no-show rate, provider preference, availability) associated with it. The number of patient calls for OA and PB appointments for a given day, and its associated attributes are uncertain. A scenario corresponds to one realization of the OA and PB appointment requests for a given day along with its attributes. Therefore, each scenario will have the following values generated:

- Number of OA calls
- Number of PB calls
- For each OA and PB call
    - Patient no-show rate
    - Service time for each phase
    - Patient's physician preference
    - Patient's availability

The attributes are randomly sampled for each patient call from a known probability distribution. The no-show rate mainly depends on the appointment delay and therefore is low for OA patients and increases as the appointment delay increases. Hence, a PB patient calling two weeks in advance for an appointment has a higher no-show rate compared to a patient calling one week in advance. The service time is the amount of time required by the nurse and the physician to treat a patient and mainly depends on the type and nature of visit. The physician preference is the patient's preferred primary care provider (PCP) and patient

availability refers to the possible times or slots during which the patient is available to see the physician.

A particular day may experience different possible scenarios. For each scenario, it is important to identify the number of OA slots to be reserved such that patients are satisfied (measured in terms of patient waiting time) and the resources are efficiently utilized (measured in terms of resource idle time and overload time). Once a scenario is generated, all the parameters associated with that scenario is deterministic. Therefore, a mixed integer linear programming model (MILP) can be used to solve each scenario. The scenario serve as inputs to the mixed integer linear programming (MILP) model discussed in Section 4.1.3. The MILP problem is solved to determine for a scenario the following:

- Number of OA and PB slots
- Position of OA and PB slots
- Number of slots to be double-booked and their position

In order to account for the uncertainties associated for a given day, multiple scenarios are generated and solved. Two different solution approaches are discussed in Section 4.2 to determine the hybrid appointment schedule for that day.

4.1.3 Mathematical Model for Two-Phase System

In this chapter, the patient scheduling problem is first formulated as a mixed integer linear programming (MILP) model for a two-phase multi-provider environment, by considering the patient's availability and provider preference. For the MILP model, the patient data is deterministic for each scenario.

Model Formulation

In this section, the notations used in the formulation of the hybrid appointment system are discussed in detail. Based on the review of literature, some model assumptions are made as follows (Cayirli and Veral, 2003; LaGanga and Lawrence, 2012):

- Patients are on-time for appointment.
- OA slots require same day appointment.
- Each phase has fixed resources (nurses and physicians).
- Patients prefer a certain physician or group of physicians.
- Overbooking of patients is allowed.
- If multiple patients are booked in a slot, then patients who called first are served first.
- Scheduling horizon is one day

Indices and Sets

$t \in T$	Set of patient types ($T = OA$ and PB)
$i \in I_t$	Set of patients of type t
$s \in S$	Set of slots in a day
$n \in N$	Set of nurses
$p \in P$	Set of physicians
$\omega \in \Omega$	Set of scenarios

Inputs

$\sigma_{i,t}$	No-show rate of patient i of type t
$\eta_{i,t}$	Nurse service time of patient i of type t
$\tau_{i,t}$	Physician service time of patient i of type t
$\rho_{i,t,p}$	1 if patient i of type t prefers physician p ; 0 otherwise
$\alpha_{i,t,s}$	1 if patient i of type t is available for slot s ; 0 otherwise
γ^N	Workload parameter of nurse[1]
γ^P	Workload parameter of physician[1]

[1]The workload parameter for a nurse or a physician is the average service time of that nurse or physician divided by the average service time of the outpatient clinic.

$NAT^B_{s,n}$	Appointment start time of nurse n in slot s
$NAT^E_{s,n}$	Appointment end time of nurse n in slot s
$PAT^B_{s,p}$	Appointment start time of physician p in slot s
$PAT^E_{s,p}$	Appointment end time of physician p in slot s
M	Very large positive number
C^{NIT}	Nurse idle time cost (\$/time)
C^{PIT}	Physician idle time cost (\$/time)
C^{NOL}	Nurse overload time cost (\$/time)
C^{POL}	Physician overload time cost (\$/time)
C^{WT}	Patient waiting time cost[2](\$/time)
C^{OC}	Opportunity cost due to an unscheduled patient[3] (\$/patient)

Decision Variables

$\delta_{i,t,s,n,p}$	1 if patient i of type t is assigned in slot s to nurse n and physician p; 0 otherwise
X_s	1 if at least one OA patient is scheduled in slot s; 0 otherwise
$NST_{i,t,s,n}$	Nurse start time of patient i of type t scheduled in slot s to nurse n
$NCT_{i,t,s,n}$	Nurse completion time of patient i of type t scheduled in slot s to nurse n
$PST_{i,t,s,p}$	Physician start time of patient i of type t scheduled in slot s to physician p
$PCT_{i,t,s,p}$	Actual physician completion time of patient i of type t scheduled in slot s to physician p
$ENST_{s,n}$	Earliest nurse start time for nurse n in slot s
$EPST_{s,p}$	Earliest physician start time for physician p in slot s

[2]The value of patient's time can be estimated by using the median hourly wage of the natives in the city where the clinic is located.

[3]The opportunity cost is the business lost by the hospital for not scheduling a patient on a given day. It can estimated by considering the revenue lost due to an unscheduled patient.

$LNCT_{s,n}$	Latest nurse completion time for nurse n in slot s
$LPCT_{s,p}$	Latest physician completion time for physician p in slot s
$ENST_{s,n}$	Earliest nurse start time for nurse n in slot s
$EPST_{s,p}$	Earliest physician start time for physician p in slot s
$LNCT_{s,n}$	Latest nurse completion time for nurse n in slot s
$LPCT_{s,p}$	Latest physician completion time for physician p in slot s
$Y_{s,n}$	1 if $LNCT_{s-1,n} < NAT^{B}_{s,n}$; 0 otherwise
$Z_{s,p}$	1 if $LPCT_{s-1,p} < PAT^{B}_{s,p}$; 0 otherwise
$ANIT_{s,n}$	Idle time of nurse n after completing service in slot s
$BNIT_{s,n}$	Idle time of nurse n before beginning service in slot s
$APIT_{s,p}$	Idle time of physician p after completing service in slot s
$BPIT_{s,p}$	Idle time of physician p before beginning service in slot s
$NWT_{i,t}$	Nurse waiting time of patient i and patient type t
$PWT_{i,t}$	Physician waiting time of patient i and patient type t
$NIT_{s,n}$	Idle time of nurse n for slot s
$NOL_{s,n}$	Overload time of nurse n for slot s
$PIT_{s,p}$	Idle time of physician p for slot s
$POL_{s,p}$	Overload time of physician p for slot s
TC	Total cost

Mathematical Formulation

The above notations are used to formulate the mathematical model. Constraints (4.1) - (4.2) ensure that a patient can be scheduled only to one slot based on the patient's physician preference and availability. Further, each slot can either accommodate a PB or an OA patient type and is given by Constraint (4.3). Note that in Constraint (4.3), i and i' are patient indices that represents two different patients, while t and t' are patient type indices that represent two different patient types. Constraint (4.4) ensures that the total number of patients scheduled to a physician in a slot is always less than or equal to two (i.e.,

double-booked).

$$\sum_{p\in P}\sum_{n\in N}\sum_{s\in S}\delta_{i,t,s,n,p} \leq 1 \qquad \forall\, t \in T,\ i \in I_t \tag{4.1}$$

$$\delta_{i,t,s,n,p} \leq \rho_{i,t,p} \times \alpha_{i,t,s} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N,\ p \in P \tag{4.2}$$

$$\sum_{n\in N}\delta_{i,t,s,n,p} + \sum_{n\in N}\delta_{i',t',s,n,p} \leq 1 \qquad \forall\, t,t' \in T,\ t \neq t',\ i \in I_t,\ i' \in I_{t'},\ s \in S,\ p \in P \tag{4.3}$$

$$\sum_{t\in T}\sum_{i\in I_t}\sum_{n\in N}\delta_{i,t,s,n,p} \leq 2 \qquad \forall\, p \in P,\ s \in S \tag{4.4}$$

In most cases, the actual service time of the patient with the nurse and physician is not known until the patient is served. Therefore, the clinic specifies the appointment start time and the appointment end time for both the nurse and the physician depending upon their workload parameter and slot duration. For example, consider a slot which begins at 9:00 am and has an appointment duration of 30 minutes. Further, the workload parameter of nurse is 1/3 and 2/3 for physician. The appointment duration of the first phase is $\frac{1}{3} \times 30 = 10$ minutes and the appointment duration of the second phase is $\frac{2}{3} \times 30 = 20$ minutes. Thus, the appointment start time and end time for the nurse is 9:00 am and 9:10 am respectively, while the appointment start time and end time for the physician is 9:10 am and 9:30 am respectively.

If an arriving patient finds the nurse (Phase I) busy, then the patient has to wait. Hence, the nurse start time for the arriving patient will not be the same as the appointment start time of that nurse in two cases: (i) when the time taken by the nurse to complete service for a patient in the previous slot exceeds the appointment start time; (ii) when the time taken by the nurse to complete service for an overbooked patient in the same slot exceeds the appointment start time. In other words, the nurse start time of a patient in Phase I will be the maximum time of the three events, namely, nurse completion time of patient scheduled in the previous slot, nurse completion time of overbooked patient in the same slot and appointment start time. Constraints (4.5) - (4.7), determines the actual nurse start time of each patient. However, if patient i is not assigned to nurse n (i.e., $\delta_{i,t,s,n,p} = 0$), then

Constraints (4.5) - (4.7) becomes redundant and Constraint (4.8) forces the start time of that patient with that nurse to be zero.

$$NST_{i,t,s,n} \geq NCT_{i',t',s-1,n} - M\big(1 - \sum_{p \in P} \delta_{i,t,s,n,p}\big) \qquad \forall\, t,t' \in T,\ i \in I_t,\ i' \in I_{t'},\ s \in S \ni s > 1,\ n \in N \qquad (4.5)$$

$$NST_{i,t,s,n} \geq NCT_{i',t,s,n} - M\big(1 - \sum_{p \in P} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i,i' \in I_t,\ i' = 1,2,..,i-1,\ s \in S,\ n \in N \qquad (4.6)$$

$$NST_{i,t,s,n} \geq NAT_n^B - M\big(1 - \sum_{p \in P} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N \qquad (4.7)$$

$$NST_{i,t,s,n} \leq M \sum_{p \in P} \delta_{i,t,s,n,p} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N \qquad (4.8)$$

On the other hand, nurse completion time for a scheduled patient is the sum of nurse service start time for that patient and the service time for the patient. However, if the patient is not scheduled, then the nurse completion time is zero. Constraints (4.9) - (4.11) determine the nurse completion time for a scheduled patient and the inclusion of a binary variable forces it to be zero for unscheduled patients.

$$NCT_{i,t,s,n} \leq NST_{i,t,s,n} + \eta_{i,t} \times (1 - \sigma_i) + M\big(1 - \sum_{p \in P} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N \qquad (4.9)$$

$$NCT_{i,t,s,n} \geq NST_{i,t,s,n} + \eta_{i,t} \times (1 - \sigma_i) - M\big(1 - \sum_{p \in P} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N \qquad (4.10)$$

$$NCT_{i,t,s,n} \leq M \sum_{p \in P} \delta_{i,t,s,n,p} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N \qquad (4.11)$$

A patient is treated by the physician (Phase II) after nurse pre-processing (Phase I). The patients can either be seen by the physician without waiting or has to wait depending upon the physician availability. Hence, the actual physician start time for a scheduled patient is the maximum time of four events, namely, physician completion time of patient scheduled in the previous slot, nurse completion time, physician completion time of overbooked patient in the same slot, or appointment start time of the physician. Constraints (4.12) - (4.15) determines the actual physician start time of each scheduled patient while Constraint (4.16) forces the physician start time to zero for an unscheduled patient.

$$PST_{i,t,s,p} \geq PCT_{i',t',s-1,p} - M\big(1 - \sum_{n \in N} \delta_{i,t,s,n,p}\big) \qquad \forall\, t,t' \in T,\ i \in I_t,\ i' \in I_{t'},\ s \in S \ni s > 1,\ p \in P \qquad (4.12)$$

$$PST_{i,t,s,p} \geq PCT_{i',t,s,p} - M\big(1 - \sum_{n \in N} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ i' = 1,2,..,i-1,\ s \in S,\ p \in P \qquad (4.13)$$

$$PST_{i,t,s,p} \geq \sum_{n \in N} NCT_{i,t,s,n} - M\big(1 - \sum_{n \in N} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \qquad (4.14)$$

$$PST_{i,t,s,p} \geq PAT^{B}_{s,p} - M\big(1 - \sum_{n \in N} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \qquad (4.15)$$

$$PST_{i,t,s,p} \leq M \sum_{n \in N} \delta_{i,t,s,n,p} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \qquad (4.16)$$

The actual physician completion time is determined using Constraints (4.17) - (4.19) and is similar to the constraints involving nurse completion time.

$$PCT_{i,t,s,p} \leq PST_{i,t,s,p} + \tau_{i,t} \times (1 - \sigma_i) + M\big(1 - \sum_{n \in N} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \qquad (4.17)$$

$$PCT_{i,t,s,p} \geq PST_{i,t,s,p} + \tau_{i,t} \times (1 - \sigma_i) - M\big(1 - \sum_{n \in N} \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \qquad (4.18)$$

$$PCT_{i,t,s,p} \leq M \sum_{n \in N} \delta_{i,t,s,n,p} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \qquad (4.19)$$

In a two phase problem, there are two possible waiting times for each patient: waiting time for nurse (Phase I) and waiting time for physician (Phase II). The waiting time for the nurse (as given in Equation 4.20) is the difference between actual nurse start time and the appointment start time of that nurse for each scheduled patient. Likewise, the waiting time for the physician (as given in Equation 4.21) is the difference between the actual physician start time and actual nurse completion time for each patient.

$$NWT_{i,t} = \sum_{s \in S} \sum_{n \in N} \big(NST_{i,t,s,n} - NAT^{B}_{n} \times \delta_{i,t,s,n,p}\big) \qquad \forall\, t \in T,\ i \in I_t \qquad (4.20)$$

$$PWT_{i,t} = \sum_{s \in S} \sum_{p \in P} PST_{i,t,s,p} - \sum_{s \in S} \sum_{n \in N} NCT_{i,t,s,n} \qquad \forall\, t \in T,\ i \in I_t \qquad (4.21)$$

A slot can be single booked or double-booked, and the nurses and physicians must serve

all the patients assigned to them in a slot before beginning service in the next slot. Further, the total expected service time of the patient(s) scheduled in a slot may be longer or shorter than the slot duration. These factors impact the latest time at which a nurse and physician completes the service in a slot, and the earliest time at which a nurse and physician can begin service in a slot. The latest nurse completion time for a slot is the completion time of the last patient scheduled to that slot, and is given by Constraint (4.22).

$$LNCT_{s,n} \geq NCT_{i,t,s,n} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ n \in N \qquad (4.22)$$

The earliest nurse start time for the first slot is the appointment time (i.e., the appointment start time of the phase for the first slot), and is given by Constraint (4.23).

$$ENST_{s,n} = NAT^{B}_{s,n} \qquad \forall\, s \in S \ni s = 1,\ n \in N \qquad (4.23)$$

The earliest nurse start time for all the other slots, depends on the latest nurse completion time of the previous slot and the appointment start time of that slot. If the appointment start time of a nurse at a slot is later than the latest nurse completion time of the nurse in the previous slot, then the earliest nurse start time at a slot will be equal to the appointment start time of the nurse for that slot. On the other hand, if the appointment start time of a nurse at a slot is earlier than the latest completion time of the nurse in the previous slot, then the earliest nurse start time at a slot will be equal to the latest nurse completion time of the previous slot.

Constraints (4.24) and (4.25) ensures that the earliest nurse start time for a slot is always greater than or equal to the maximum of the appointment start time of a nurse in that slot and latest completion time of the nurse in the previous slot, respectively. Further, Constraints (4.26) is active (i.e., $ENST_{s,n} \leq NAT^{B}_{s,n}$) only when the binary variable, $Y_{s,n}$, is 1 and Constraint (4.27) is active (i.e., $ENST_{s,n} \leq LNCT_{s-1,n}$) when the binary variable, $Y_{s,n}$, is 0. To ensure feasibility of (4.24) and (4.25), the binary variable, $Y_{s,n}$, will be 1 if appointment start time of a nurse for a slot is later than the latest completion time of the

nurse in the previous slot and 0 otherwise. Thus, Constraints (4.24) and (4.26) will force $ENST_{s,n}$ to be exactly equal to $NAT^B_{s,n}$, when $Y_{s,n}$ = 1 (i.e., $NAT^B_{s,n} > LNCT_{s-1,n}$) and Constraints (4.25) and (4.27) will force $ENST_{s,n}$ to be exactly equal to $LNCT_{s-1,n}$, when $Y_{s,n} = 0$.

$$ENST_{s,n} \geq NAT^B_{s,n} \qquad \forall\, s \in S \ni s > 1,\ n \in N \tag{4.24}$$

$$ENST_{s,n} \geq LNCT_{s-1,n} \qquad \forall\, s \in S \ni s > 1,\ n \in N \tag{4.25}$$

$$ENST_{s,n} \leq NAT^B_{s,n} + M(1 - Y_{s,n}) \qquad \forall\, s \in S \ni s > 1,\ n \in N \tag{4.26}$$

$$ENST_{s,n} \leq LNCT_{s-1,n} + MY_{s,n} \qquad \forall\, s \in S \ni s > 1,\ n \in N \tag{4.27}$$

The latest physician completion time is given by Constraint (4.28), and is similar to Constraint (4.22) associated with the latest nurse completion time. The earliest physician start times are determined using Constraints (4.29) - (4.33), and are similar to Constraints (4.23) - (4.27) associated with the earliest nurse start times.

$$LPCT_{s,p} \geq PCT_{i,t,s,p} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P \tag{4.28}$$

$$EPST_{s,p} = PAT^B_{s,p} \qquad \forall\, s \in S \ni s = 1,\ p \in P \tag{4.29}$$

$$EPST_{s,p} \geq PAT^B_{s,p} \qquad \forall\, s \in S \ni s > 1,\ p \in P \tag{4.30}$$

$$EPST_{s,p} \geq LPCT_{s-1,p} \qquad \forall\, s \in S \ni s > 1,\ p \in P \tag{4.31}$$

$$EPST_{s,p} \leq PAT^B_{s,p} + M(1 - Z_{s,p}) \qquad \forall\, s \in S \ni s > 1,\ p \in P \tag{4.32}$$

$$EPST_{s,p} \leq LPCT_{s-1,p} + MZ_{s,p} \qquad \forall\, s \in S \ni s > 1,\ p \in P \tag{4.33}$$

The nurse can remain idle in two situations: (i) before beginning the service and (ii) after completing the service. At a given slot, if a nurse does not begin service at the earliest possible start time, then that nurse is idle before beginning the service, and is given by Constraint (4.34). On the other hand, if a nurse serves all the patient assigned to a slot before

the appointment end time of that nurse in that slot, then the nurse is idle after completing the service. However, if the latest nurse completion time in a slot exceeds the appointment end time of the nurse in that slot, then the nurse is overburdened in that slot. The nurse idle time after service completion and the nurse overload time for each slot is given by Constraint (4.35).

$$BNIT_{s,n} \geq LNCT_{s,n} - ENST_{s,n} - \sum_{t \in T} \sum_{i \in I_t} \sum_{p \in P} \eta_{i,t} \times (1 - \sigma_i) \times \delta_{i,t,s,n,p} \qquad \forall\, s \in S,\ n \in N \qquad (4.34)$$

$$ANIT_{s,n} - NOL_{s,n} = NAT^{E}_{s,n} - LNCT_{s,n} \qquad \forall\, s \in S,\ n \in N \qquad (4.35)$$

Similarly, the physician can also remain idle before beginning the service and after completing the service. If the physician has to wait for the nurse to complete the pre-screening procedure, then the physician is idle before beginning the service, and is given by Constraint (4.36). If the physician serves all the patients scheduled in a slot before the physician appointment end time of that slot, then the physician is idle after completing the service. However, if time required to serve all the patients scheduled in a slot exceeds the appointment end time of that slot, then the physician is overloaded in that slot. Constraint (4.37) determines the idle time of physicians after completing service and the physician overload time at a slot.

$$BPIT_{s,p} \geq LPCT_{s,p} - EPST_{s,p} - \sum_{t \in T} \sum_{i \in I_t} \sum_{n \in N} \tau_{i,t} \times (1 - \sigma_i) \times \delta_{i,t,s,n,p} \qquad \forall\, s \in S,\ p \in P \qquad (4.36)$$

$$APIT_{s,p} - POL_{s,p} = PAT^{E}_{s,p} - LPCT_{s,p} \qquad \forall\, s \in S,\ p \in P \qquad (4.37)$$

The non-negativity and binary restrictions on the decision variables is ensured using Constraints (4.38) and (4.39), respectively.

$$\begin{aligned} &NST_{i,t,s,n}, NCT_{i,t,s,n}, PST_{i,t,s,p}, PCT_{i,t,s,p}, LNCT_{s,n}, \\ &NIT_{s,n}, BNIT_{s,n}, ANIT_{s,n}, NOL_{s,n}, LPCT_{s,p}, PIT_{s,p}, \\ &BPIT_{s,p}, APIT_{s,p}, POL_{s,p}, NWT_{i,t}, PWT_{i,t} \geq 0 \end{aligned} \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P,\ n \in N \qquad (4.38)$$

$$\delta_{i,t,s,n,p},\ Y_{s,n},\ Z_{s,p} \in 0,1 \qquad \forall\, t \in T,\ i \in I_t,\ s \in S,\ p \in P,\ n \in N \tag{4.39}$$

Objective Function

The objective function (Equation 4.40) is the sum of physician and nurse idle time cost, patient waiting time cost and the opportunity cost. The total idle time cost is the product of cost of being idle per unit time and the total idle time. Similarly, the total waiting time cost is the product of cost of waiting per unit time and the total waiting time. The opportunity cost is the product of cost of not scheduling a patient and the total number of unscheduled patients.

$$\begin{aligned} \text{Minimize} \quad Z = \quad & C^{PIT}\Big[\sum_{s\in S}\sum_{p\in P} BPIT_{s,p} + \sum_{s\in S}\sum_{p\in P} APIT_{s,p}\Big] + C^{NIT}\Big[\sum_{s\in S}\sum_{n\in N} BNIT_{s,n} + \sum_{s\in S}\sum_{n\in N} ANIT_{s,n}\Big] + \\ & C^{POL}\Big[\sum_{s\in S}\sum_{p\in P} POL_{s,p}\Big] + C^{NOL}\Big[\sum_{s\in S}\sum_{n\in N} NOL_{s,n}\Big] + \\ & C^{WT}\Big[\sum_{t\in T}\sum_{i\in I_t}(NWT_{i,t} + PWT_{i,t})\Big] + C^{OC}\Big[\sum_{t\in T}\sum_{i\in I_t}\sum_{s\in S}\sum_{n\in N}\sum_{p\in P}(1-\delta_{i,t,s,n,p})\Big] \end{aligned} \tag{4.40}$$

The size (i.e., number of constraints, number of binary variables, and number of continuous variables) of the MILP model depends on the problem parameters, namely, number of OA patient calls, number of PB patient calls, number of slots, number of physicians, and number of nurses. For instance, consider a situation in which there are 7 OA patient calls, 26 PB patient calls, 12 slots, 2 physicians and 2 nurses. The MILP model representing the above situation has 84,646 linear constraints, 3,523 continuous variables and 1,661 binary variables.

4.2 Determining the Hybrid Appointment Schedule

The MILP model provides the optimal schedule for a given set of input parameters such that the total cost is minimized. Using the optimal schedule, the following output measures are determined for the hybrid appointment system.

- Number of OA and PB slots
- Position of OA and PB slots

- Number of slots to be double-booked
- Position of the slots to be double-booked

A scenario for a typical day (discussed in Section 4.1.2) is used as an input to the MILP model. Most outpatient departments will have to handle various stochastic parameters, such as the total number of patient calls for an appointment for a given day, the patient no-show rates, patient service times, and patient availability to make good decisions. In order to account for the stochastic parameters, the MILP model is solved for several scenarios for a particular day. Each scenario will produce an optimal schedule and the set of output measures. The solutions of all the scenarios obtained from the MILP model are used to generate the final configuration of the hybrid appointment system (i.e., the set of output measures) for that day. We will discuss two solution approaches for determining the final configuration.

4.2.1 Deterministic Equivalent Program (DEP)

If there are only relatively small number of scenarios, then the problem can be modeled as a large mixed integer linear programming (MILP) problem. This formulation is often called the Deterministic Equivalent Program of a stochastic problem (Birge and Louveaux, 2011). The deterministic equivalent model includes similar set of constraints for each realization of the random data or scenario. In other words, the constraint set is similar for each scenario and the objective function is written as the sum of the product of probability of occurrence of each scenario and the total cost for each scenario. The objective function is to minimize the expected total cost. Therefore, the problem size increases linearly with the number of scenarios.

The MILP problem discussed in Section 4.1.3 can be modeled and solved as a deterministic equivalent program. All the variables and parameters used in the mathematical formulation will have an additional index as a superscript to denote a particular scenario ($\omega \in \Omega$). For example, $NST^{\omega}_{i,t,s,n}$ denotes the nurse start time of patient i of type t by nurse n in slot s under scenario ω. Also, TC^{ω} denotes the total cost for scenario ω.

For the problem under consideration, we must determine the total number of OA slots to be reserved, such that the expected total cost across all the scenarios are minimized. Therefore, the total number of slots reserved for OA patients must be the same across all the scenarios. In order to formulate the DEP, Constraints (4.1) - (4.39) must be repeated for each scenario. Further, to ensure that the OA slots and their position are the same across all the scenarios, it is necessary to add additional constraints linking all the scenarios. A binary variable X_s is used as a linking variable. A slot is considered as an OA slot, if the value of X_s is one and is considered as a PB slot if its value is zero. Constraint (4.41) serves as a lower bound and forces X_s to be 1, when at least one of the OA patients is scheduled in slot s. On the other hand, Constraint (4.42) serves as an upper bound and ensures that X_s is equal to 0, when none of the OA patients is scheduled in slot s. Therefore, if the solution indicates the value of X_s to be 1, then slot s must be reserved as an OA slot across all the scenarios, and if the solution indicates the value of X_s to be 0, then slot s must be reserved as a PB slot across all scenarios

$$X_s \geq \delta^{\omega}_{i,t,s,n,p} \qquad t \in OA,\ i \in I_t,\ s \in S,\ p \in P,\ n \in N,\ \omega \in \Omega \qquad (4.41)$$

$$X_s \leq \sum_{t \in OA} \sum_{i \in I_t} \sum_{n \in N} \sum_{p \in P} \delta^{\omega}_{i,t,s,n,p} \qquad s \in S,\ \omega \in \Omega \qquad (4.42)$$

Drawbacks of the Deterministic Equivalent Model

When the number of slots, number of resources (nurse and physician), or number of scenarios increase, the number of binary variables also increase, resulting in longer solution times. The author verified and solved the deterministic equivalent program (DEP) for a small setting as part of his Master's paper (Srinivas, 2015). The author considered a clinic setting which has 1 nurse, 1 physician, 4 slots. Further, two scenarios were considered. In scenario 1, there were 5 OA and 3 PB patients calls for appointment, and in scenario 2 there were 3 OA and 5 PB patients calls for appointment. The DEP had 2,326 linear constraints and 454 variables (370 continuous and 84 binary). The model was solved in a computer with 8 GB RAM using i5 processor in 7.3 minutes, with an optimality gap of 0.01% (i.e.,

the difference between the best known solution and the best bound is 0.01%).

As part of this dissertation, we attempted to solve the DEP with 5 OA patient calls, 15 PB patient calls, 8 slots, 2 physicians, 2 nurses, and 5 different scenarios with the same parameters. In other words, only the service times and the patient no-show values will change across the scenarios. For such a situation, each scenario has 22,041 linear constraints, 1,513 continuous variables and 688 binary variables. Therefore, for 5 scenarios, there will be 110,205 linear constraints (22,041 linear constraints/scenario $\times$ 5 scenarios), 7,565 continuous variables (1,513 continuous variables/scenario $\times$ 5 scenarios), and 3,440 binary variables (688 binary variables/scenario $\times$ 5 scenarios). In addition, the deterministic equivalent program must include additional variables and constraints to link all the scenarios (Equations 4.41 and 4.42). For this situation, there will be 840 additional constraints and 8 additional binary variables to link all the scenarios.

Hence, the deterministic equivalent program representing the above situation will have 111,045 linear constraints, 7,565 continuous variables and 3448 binary variables. When the model was executed in CPLEX, it indicated that the solver ran out of memory. Therefore, it was not possible to obtain a solution even for a small clinic setting. Hence, a better solution approach is necessary.

4.2.2 Monte Carlo Scenario Analysis

In the Monte Carlo scenario analysis, the MILP model is executed independently for each scenario. Hence, the computational effort required to solve each MILP would be much less compared to the Deterministic Equivalent approach and more realistic clinic settings can be evaluated. The following values are determined for each scenario by the MILP model:

- Number of OA and PB slots
- Position (Timings) of OA and PB slots
- Number of slots to be double-booked and their positions

Note that due to the absence of a linking variable in the MILP model, the number and

position of each slot-type need not be the same across all scenarios. Hence, heuristics are developed to select a common schedule that would be best across all the scenarios.

4.2.3 Heuristic Approaches

In order to ensure adaptability by hospitals, it is necessary to determine one best appointment schedule configuration that performs well under uncertainty. Therefore, two heuristic approaches are developed to analyze the results of all the scenarios and select the best schedule that determines the number of slots reserved for open access and pre-book, position of pre-book and open access slots, number of slots double-booked and their position in the schedule.

4.2.3.1 Heuristic 1: Frequency Method

Heuristic 1 uses the most frequently appearing value (i.e., votes) across all scenarios to determine the number and position of open access, pre-booked, and double-booked slots in the final configuration.

- *Number of open access and pre-book slots*: The total number of open access slots to be reserved for a physician will be equal to the number of slots that are reserved as open access in most of the scenarios for that physician. Similarly, the total number of pre-book slots to be reserved for a physician will be equal to the number of slots that are reserved as pre-book in most of the scenarios for that physician.

- *Number of double-booked slots*: The total number of open access and pre-book slots to be double-booked will be equal to the total number of open access and pre-book slots that are double-booked most often across all the scenarios, respectively.

- *Position of open access and pre-book slots*: Each slot is analyzed independently for all the scenarios, and the position of open access slot is determined based on the number of times a particular slot is selected as an open access slot across all the scenarios. Therefore, the appointment slots that are most frequently considered as

open access across all the scenarios are chosen as the position of open access slots. The remaining slot positions are assigned as pre-book slots.

- *Position of double-booked slots*: The position of the open access slot or pre-book slot to be double-booked is determined based on the number of times a particular slot type is double-booked across all the scenarios, and is similar to the procedure for determining the position of the open access slot.

4.2.3.2 Heuristic 2: Averaging Method

Heuristic 2 uses the average across all the scenarios to determine the number and position of slots reserved for each appointment type in the final configuration.

- *Number of open access and pre-book slots*: The total number of pre-book slots to be reserved for a physician will be equal to the average of the number of slots that are reserved as pre-book across all the scenarios rounded to the nearest integer for that physician. Similarly, the total number of open access slots to be reserved for a physician will be equal to the average of the number of slots that are reserved as open access across all the scenarios rounded to the nearest integer for that physician. Note that when rounding to the nearest integers, it is necessary to ensure that the sum of the number of open access and pre-book slots is equal to the total number of slots.

- *Number of double-booked slots*: The total number of open access and pre-book slots to be double-booked will be equal to the average of the number of open access and pre-book slots that are double-booked across all the scenarios, respectively.

- *Position of open access and pre-book slots*: The position of the open access slots are determined using weighted average of each slot, where the weights are the probabilities of an event occurring. For each slot, the weighted average of that slot being selected as an open access slot is calculated using Equation 4.43. Then, the position of the open access slots are selected based on their weighted average values

(higher the better). The remaining slot positions are then considered as pre-book slots.

$$\text{Weighted Average for Slot } s \ (\mu_s) = \sum_i x_{is} p_{is} \tag{4.43}$$

where, x_{is} is outcome i for slot s being open access

p_{is} is the probability of outcome i for slot s

A slot can either be open access slot or a pre-book slot. Therefore, if $x_{is} = 1$, then the slot s is open access and if $x_{is} = 0$, then slot s is pre-book. Hence, Equation 4.43 reduces to Equation 4.44.

$$\mu_s = 1 \times p(\text{selecting slot } s \text{ as open access}) + 0 \times p(\text{not selecting slot } s \text{ as open access}) \tag{4.44}$$

It is evident from Equation 4.44 that the weighted average of selecting slot s is always equal to the probability of selecting slot s as open access.

- *Position of double-booked slots*: The position of the open access slot or pre-book slot to be double-booked is determined based on the weighted average of that slot being double-booked, and is similar to the procedure for determining the position of the open access slot.

4.2.4 Implementation of the Hybrid Appointment Policy

The final configurations obtained by using Heuristic 1 and Heuristic 2 are back-tested across all the scenarios. For each heuristic, the back-testing algorithm fixes the schedule configuration obtained using that heuristic and mimics the scheduling process for all the scenarios, and determines the performance measures, namely, average waiting time, average resource idle time, average resource overload time, and average number of unscheduled patients. The hospital administrators can choose the best configuration (i.e., number of open access and pre-book slot, their timings, and the position of slots to double-book for a typical day) by comparing the values of the performance measures obtained for each heuristic.

The procedure is repeated for each working day of the week (Monday - Friday). Each working day is analyzed separately because the patient call volumes are generally higher at the beginning of the week. If a patient calling for an appointment requests an open access appointment, then the scheduling manager schedules the patient to one of the open access slots on the day of the patient call. On the other hand, if a patient calling for an appointment requests a pre-book appointment, then the scheduling manager schedules the patient to one of the pre-book slots on any day starting with the next day of the patient call. Further, the scheduling manager also knows the number of slots that he can double-book based on the final configuration. Therefore, by using this procedure, the scheduling manager will be able to construct an hybrid appointment schedule that will improve both resource utilization and patient satisfaction.

4.3 Case Study

4.3.1 Case Study Background

The proposed methodology is demonstrated using a case study with real data. The data was obtained from a Family Medicine clinic in Pennsylvania and included the de-identified medical records of patient visits from July 2013 to July 2015. A specific day of the week was selected, and the corresponding data for the last 2 years for that day of the week was used to determine the model parameters. The clinic accepts both pre-booking and open access appointments, and serves the patients in two phases. An arriving patient is first seen by the nurse and then by the physician. Further, the clinic has two FTE nurse and two FTE physicians on-duty for the day under consideration. Each resource (nurse and physician) works for 12 slots per day.

4.3.2 Data Collection and Parameter Estimation

On analyzing the total patient calls for the last two years, it was found that it follows a Poission distribution with a mean of 30 calls per day. Further, 15%-25% of the patient

calls were open access (same day) appointment requests. Out of the total patient calls, 25% preferred to see Physician 1, 25% preferred to see Physician 2 and the remaining 50% had no preference (i.e., willing to see either Physician 1 or Physician 2). The clinic provides a 30-minute time slot to each calling patient, and the nurse is expected to serve the patient in the first 10 minutes and the physician is expected to serve the patients in the remaining 20 minutes. However, the actual service time may vary, and the resource may complete their service earlier or later than the appointment end time. The service time for phase 1 (i.e., nurse) follows a uniform distribution over the interval (5,15) and the service time for phase 2 also follows a uniform distribution over the interval (15, 25). In addition, the average no-show rate of open access patients is 5% and the average no-show rate of a pre-booked patients is 20%. A total of 20 scenarios is evaluated by varying the parameters. All the scenarios are assumed to be equally likely.

Based on the city in which the hospital is located, the median salary of full time equivalent (FTE) physicians is about $192,000/year, the median salary of FTE nurses is about $67,500/year, and the median salary of the patients (city population) is about $58,500/year. If physicians and nurses work for 8 hours a day and 5 days a week, then the hourly rate can be computed by dividing the annual salary by 2080 hours (8 hours/day $\times$ 5 days/week $\times$ 52 weeks/year = 2080 hours/year). Therefore, the hourly rate of FTE physicians, nurses and patients are approximately $92, $32, and $28. Hence, if the physician is idle for a minute, then the hospital loses $1.53 (i.e., 92/60), and if the nurse is idle for a minute, then the hospital loses $0.53 for each minute of nurse idle time. Similarly, the patient loses $0.47 for each minute of waiting. Based on the interaction with the staff at the hospital, it is determined that the average net revenue of treating a patient is $40. Generally, the patients are expected to be served within the 30-minute time slot. However, if a resource at a slot serves beyond the appointment end time of the slot, then it may lead to resource burnout, emotional stress, medical errors, decrease in productivity, hospital acquired infections, and early retirement. In this paper, the cost of resource overload per minute is assumed to be 1.5 times the resource idle time per minute. In other words, each

minute of nurse overload is \$0.80 (i.e., 0.53×1.5) and each minute of physician overload is \$2.30 (i.e., 1.53×1.5).

A summary of the model parameters used to generate the input values for each scenario is shown in Table 4.1.

Table 4.1: Summary of model parameters

Parameter	**Value**
Total patient calls for appointment	Poisson with rate $\lambda = 30$
Proportion of calls requesting open access	15%, 20%, 25% with probability 0.3, 0.4, 0.3
Average no-show rate of open access patient	5%
Average no-show rate of pre-booked patient	20%
Number of FTE nurses ($\|N\|$)	2
Number of FTE physicians ($\|P\|$)	2
Proportion of patients preferring Physician 1	25%
Proportion of patients preferring Physician 2	25%
Proportion of patients without any preference	50%
Service time in minutes for Phase 1 (Nurses)	Uniform $\sim$ (5,15)
Service time in minutes for Phase 2 (Physicians)	Uniform $\sim$ (15,25)
Total number of slots in a day ($\|S\|$)	12 slots
Slot duration (D)	30 minutes
Costs/hour ($C^{WT}, C^{NIT}, C^{NOL}, C^{PIT}, C^{POL}$)	\$ $28, 32, 48, 92, 138$
Revenue lost/unscheduled patient (C^{OC})	\$40
Total number of scenarios (Ω)	20

Each scenario for the outpatient clinic corresponds to one realization of the random events and will have the following values generated:

- Number of open access calls
- Number of pre-book calls
- For each open access and pre-book call

- Patient no-show rate
- Service time for each phase
- Patient's physician preference
- Patient's availability

The sample values for each scenario are drawn using the probability distributions given in Table 4.1. For example, the first scenario had 33 patient calls for appointment (7 open access calls and 26 pre-book calls). The problem size of such a scenario is as follows:

- Number of constraints: 84,646
- Number of binary variables: 1,661
- Number of continuous variables: 3,523

The scenario values are then input to the MILP model. The MILP model is coded in Microsoft Visual C++ 6.0 and solved on a computer (1.60 GHz, 8.00 GB RAM) using Concert Technology framework of ILOG CPLEX 12.4, with an optimality gap of 10%. The model was solved in 10.4 minutes. Similarly, the model is independently solved for all the 20 scenarios with the data for each scenario to account for the variability in the stochastic parameters. The solution time for each scenario is approximately 10 minutes. Hence, the time taken to solve 20 scenarios is approximately 200 minutes.

The results of all the independent runs are analyzed using the heuristic approach discussed in Section 4.2.2 to determine the number of open access slots, number of pre-book slots, position of open access slots, position of pre-book slots, number of slots to double-book, and the position of the slots to double-book.

4.3.3 Case Study Results

4.3.3.1 Determining Number of Slots for Each Appointment Type

Table 4.2 presents the optimal number of open access, pre-book, and double-book slots for each physician across all the 20 scenarios. In addition, it also provides the most frequent value and the average value across all the scenarios. For example, for Physician 1, the most frequently occurring optimal value for open access slots is 2 (9 out of the 20 scenarios) and the average number of slots reserved for open access appointments is 2.65. Hence, based on Heuristic 1 (Frequency method), Physician 1 will have 2 open access slots, and based on Heuristic 2 (Averaging method), Physician 1 will have 3 (i.e., 2.65 rounded to nearest integer) open access slots. Similarly, the number of slots reserved for each appointment type for each physician can be determined using Heuristics 1 and 2.

The final configuration for Physician 1 based on Heuristic 1 will have 2 open access slots, 10 pre-book slots, and 3 of the 10 pre-book slots can be double-booked. Similarly, for Physician 2, 3 open access, 9 pre-book slots and 3 out of the 9 pre-book slots can be double-booked. The final configuration based on Heuristic 2 will have 3 open access slots, 9 pre-book slots, and 4 of the 10 pre-book slots can be double-booked for both Physician 1 and Physician 2. Further, based on the heuristics, none of the open access slots can be double-booked for both the physicians in the final configuration.

Table 4.2: Number of slots for each appointment type across 20 scenarios

Scenario	Physician 1				Physician 2			
	No. of OA slots	**No. of PB slots**	**No. of OA slots double-booked**	**No. of PB slots double-booked**	**No. of OA slots**	**No. of PB slots**	**No. of OA slots double-booked**	**No. of PB slots double-booked**
Scenario 1	4	8	0	3	2	10	1	4
Scenario 2	3	9	0	3	2	9	1	5
Scenario 3	1	11	0	4	4	8	0	3
Scenario 4	2	10	1	3	4	9	0	4
Scenario 5	2	10	0	5	1	10	1	3
Scenario 6	2	10	0	3	2	10	0	3
Scenario 7	3	9	1	2	3	9	1	4
Scenario 8	4	8	0	3	3	9	0	5
Scenario 9	4	8	0	3	3	9	0	3
Scenario 10	2	10	1	6	3	9	0	3
Scenario 11	2	10	0	5	3	9	0	5
Scenario 12	3	9	0	3	3	9	0	6
Scenario 13	2	10	0	5	3	9	0	4
Scenario 14	3	9	0	3	4	8	1	4
Scenario 15	2	10	1	4	3	9	0	3
Scenario 16	3	9	0	3	2	9	0	5
Scenario 17	2	10	0	3	4	8	0	5
Scenario 18	4	8	0	3	3	9	0	3
Scenario 19	3	9	0	3	3	9	0	6
Scenario 20	2	10	1	4	3	9	1	3
Most Frequent	**2**	**10**	**0**	**3**	**3**	**9**	**0**	**3**
Average	**2.65**	**9.35**	**0.25**	**3.55**	**2.9**	**9**	**0.3**	**4.05**

4.3.3.2 Determining Position of Open Access and Pre-book Slots

Figure 4.2 and Table 4.3 presents the number of scenarios in which a particular slot is reserved as open access slot for each Physician. The open access appointment requests occur on the same day of the patient's visit. Therefore, it would be impractical to allocate the first few slots as open access slot because the patient calling at the beginning of the day will not be able to visit the clinic during those slots. Therefore, the first four slots are always reserved as pre-book slot for both the Physicians, and is given as an input to the MILP model. Thus, in Figure 4.2 and Table 4.3, Slot 1 through Slot 4 are not considered because they are not reserved as open access slot in any of the scenarios.

Based on Heuristic 1, Physician 1 has 2 OA slots and Physician 2 has 3 OA slots. Hence, it is necessary to choose the two most frequent slot positions that are occurring as open access slots for Physician 1 and three most frequent slot positions that are occurring as open access slot for Physician 2. It is clear from Figure 4.2 and Table 4.3 that Slots 6 and 8 will be the open access slot positions for Physician 1 and Slots 9, 10, and 11 will be the open access slot positions for Physician 2 under Heuristic 1. The remaining slot positions are reserved as pre-book slots for Physicians 1 and 2.

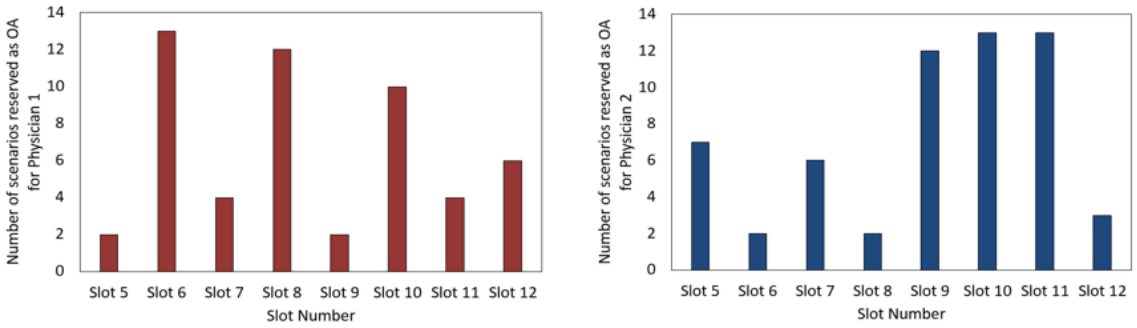


Figure 4.2: Number of scenarios in which a slot is reserved as OA slot for Physician 1 and Physician 2

Table 4.3: Number of scenarios in which a slot is reserved as open access

	Physician 1		Physician 2	
Slot Position	**No. of scenarios reserved as OA**	**Weighted Average***	**No. of scenarios reserved as OA**	**Weighted Average***
Slot 5	2	0.10	7	0.35
Slot 6	13	0.65	2	0.10
Slot 7	4	0.20	6	0.30
Slot 8	12	0.60	2	0.10
Slot 9	2	0.10	12	0.60
Slot 10	10	0.50	13	0.65
Slot 11	4	0.20	13	0.65
Slot 12	6	0.30	3	0.15

* All scenarios are assumed to be equally likely.

Based on Heuristic 2, it is necessary to choose the three slot positions that have a high weighted average score for both Physician 1 and Physician 2. The weighted average is calculated using Equation 4.44. For example, for Physician 1, the probability of selecting Slot 5 as open access is $\frac{2}{20}$ (i.e., slot 5 is selected as open access in 2 out of the 20 scenarios), and the probability of not selecting slot 5 as open access is $\frac{18}{20}$. Further, all the scenarios are assumed to be equally likely. Substituting the values of the probability in Equation 4.44, we get the weighted average for selecting Slot 5 as open access as 0.10. Therefore, based on the weighted average values in Table 4.3, Slots 6, 8 and 10 will be the open access slot positions for Physician 1 and Slots 9, 10, and 11 will be the open access slot positions for Physician 2 under Heuristic 2. The remaining slot positions are reserved as pre-book slots for Physicians 1 and 2.

4.3.3.3 Determining the Position of Slots to Double-Book

Based on the analysis in Section 4.3.3.1 (Table 4.2), none of the open access slots should be double-booked under Heuristics 1 and 2. For the pre-book slots, three slots should be double-booked for each physician under Heuristic 1 and four slots should be double-booked for each physician under Heuristic 2. Figure 4.3 and Table 4.4 shows the number of scenarios in which a pre-book slot is double-booked and the weighted average value for each slot position. Note that 3 pre-book slots have to be chosen for double-booking under Heuristic 1 and 4 under Heuristic 2 for both physicians. Slots 1, 3 and 4 are the top three slot positions that are frequently double-booked for both the physicians and are reserved as the positions for double-booking under Heuristic 1. The first four slot positions has the highest weighted average value for both the physicians and are reserved as the positions for double-booking under Heuristic 2.

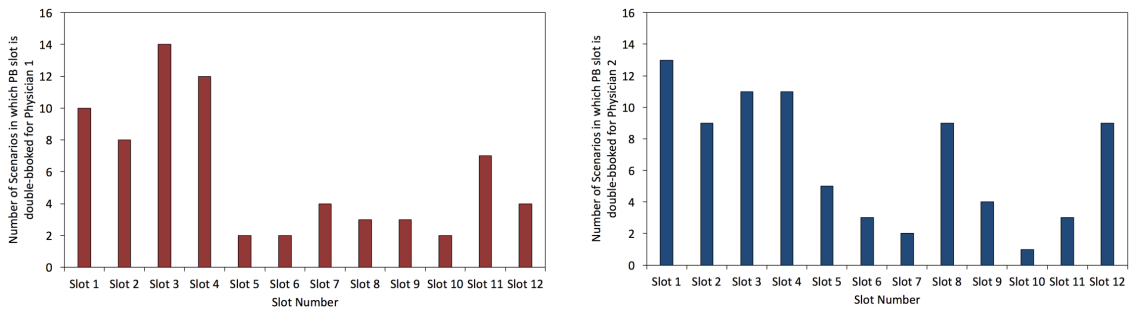


Figure 4.3: Number of scenarios in which a pre-book slot is double-booked for Physician 1 and Physician 2

Further, the analysis of the optimal schedule generated from each scenario indicated that a slot, which is double-booked always had at least one patient with a high no-show rate. In other words, it is optimal to schedule at least one patient with a high probability of not visiting the clinic to the double-book slot.

Table 4.4: Number of scenarios in which a pre-book slot is double-booked

	Physician 1		Physician 2	
Slot Position	**No. of scenarios with double-booked PB slot**	**Weighted Average**	**No. of scenarios with double-booked PB slot**	**Weighted Average**
Slot 1	10	0.50	13	0.65
Slot 2	8	0.40	9	0.45
Slot 3	14	0.70	11	0.55
Slot 4	12	0.60	11	0.55
Slot 5	2	0.10	5	0.25
Slot 6	2	0.10	3	0.15
Slot 7	4	0.20	2	0.1
Slot 8	3	0.15	9	0.45
Slot 9	3	0.15	4	0.2
Slot 10	2	0.10	1	0.05
Slot 11	7	0.35	3	0.15
Slot 12	4	0.20	9	0.45

4.3.3.4 Final Configuration of the Hybrid Appointment System

There will be one final configuration based on each heuristic. The final configuration for Physician 1 and Physician 2 based on Heuristic 1 and Heuristic 2 are shown in Figures 4.4 and 4.5.

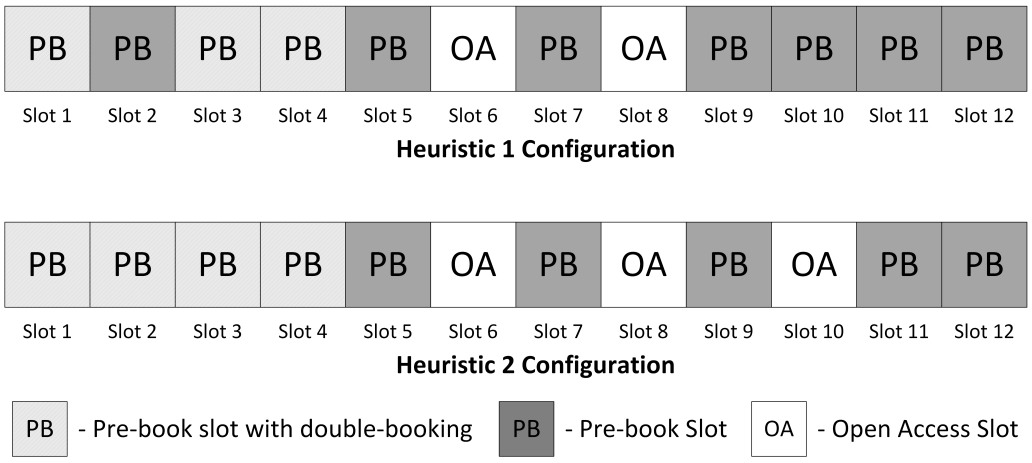


Figure 4.4: Final configuration of hybrid appointment system for Physician 1 under Heuristic 1 and Heuristic 2

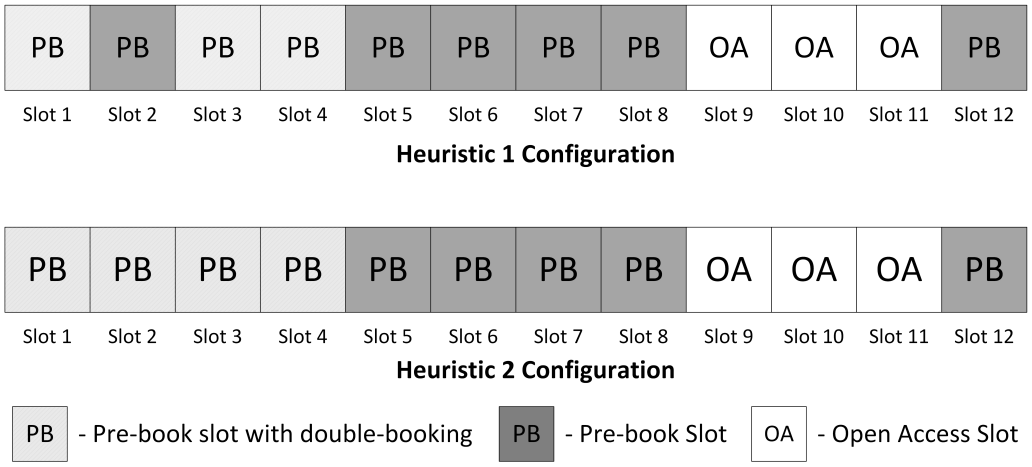


Figure 4.5: Final configuration of hybrid appointment system for Physician 2 under Heuristic 1 and Heuristic 2

The configuration of the hybrid appointment system is practical because of the following reasons:

- **Slots are double-booked at the beginning**: Double-booking increases the resource utilization. Therefore, if the slots are double-booked in the beginning, then the physician has steady flow of patients to begin the day.
- **Double-book with at least one high probability no-show patient**: If a slot is double-booked and if both the patients show-up for their appointment, then the waiting time of the upcoming patients in the schedule will increase. Further, the resource will be overloaded. Double-booking with at least one high probability no-show patient reduces the probability of both the patients showing up for an appointment.
- **Open access slots are reserved in the second half of the schedule**: A patient requesting open access appointment (i.e., same day appointment request), may not be able to visit the clinic at the beginning of the day. Further, some open access appointment requests may also occur later in the day. Therefore, having open access slots in the second half of the schedule appears to be practical.

4.3.4 Backtesting of the Final Configuration

The final configurations obtained from the two heuristics are backtested with the past data to evaluate their performance, namely, average patient waiting time, average resource overtime, average resource idle time, and average number of unscheduled patients. The total number of patients who can be scheduled to a physician is fixed based on the configuration. For example, Physician 1 can accommodate 15 patients (2 open access and 13 pre-book patients). Therefore, a patient may not get an appointment for a particular day (i.e., unscheduled patient) when the schedule is fully booked or the patient's availability or preference cannot be accommodated with the current configuration. A simulation model is developed to mimic the sequential scheduling process. In addition, the simulation model also takes into account the insights obtained from the MILP model by double-booking with at least one high no-show patient. The parameters of the 20 scenarios that were used as inputs to the optimization model (i.e., patient calls, availability, preference, no-show rate, service times) are given as input to the simulation model for computing the performance measures for each of the 20 scenarios. However, unlike the mathematical model, the schedule configuration of the simulation model is fixed based on the heuristic begin evaluated. The performance of both the heuristics is shown in Table 4.5.

Table 4.5: Comparison of performance measures of Heuristic 1 and Heuristic 2

Performance Measure	**Heuristic 1 (Frequency Method)**	**Heuristic 2 (Averaging Method)**
Average nurse idle time	20.55	18.63
Average nurse overtime	19.80	22.23
Average physician idle time	42.50	39.43
Average physician overtime	50.66	67.47
Average patient waiting time	6.40	8.05
Average no. of unscheduled patients	4.29	3.14

Heuristic 1 performs better with respect to average nurse overtime, average physician

overtime, average patient waiting time, while Heuristic 2 is better with respect to average nurse idle time, average physician idle time and average no. of patients rejected. The overall cost associated with the schedule generated using each heuristic can be computed using the current cost settings in Table 4.1. Heuristic 1 will have an average total cost of $677.50 and Heuristic 2 will have an average total cost of $729.21. Therefore, Heuristic 1 is better than Heuristic 2 in terms of the overall cost. However, if the hospital administration is interested in reducing the number of patients rejected and the average idle time of the resources, then they may choose the configuration given by Heuristic 2.

4.4 Generalized Formulation for a Multi-Phase Multi-Provider System

Sections 4.1 and 4.2 discussed models and methods for a clinic with only two phases/stations (nurse and physician). In practice, it is possible to experience more than two phases. For instance, an arriving patient might visit four phases (stations), namely, check-in, see nurse, see physician, and check-out as shown in Figure 4.6. Therefore, in this section, the mathematical model proposed in Section 4.1 is extended to develop a generalized mathematical model for multi-phase, multi-provider, hybrid appointment system. The generalized model can handle any number of phases and resources.

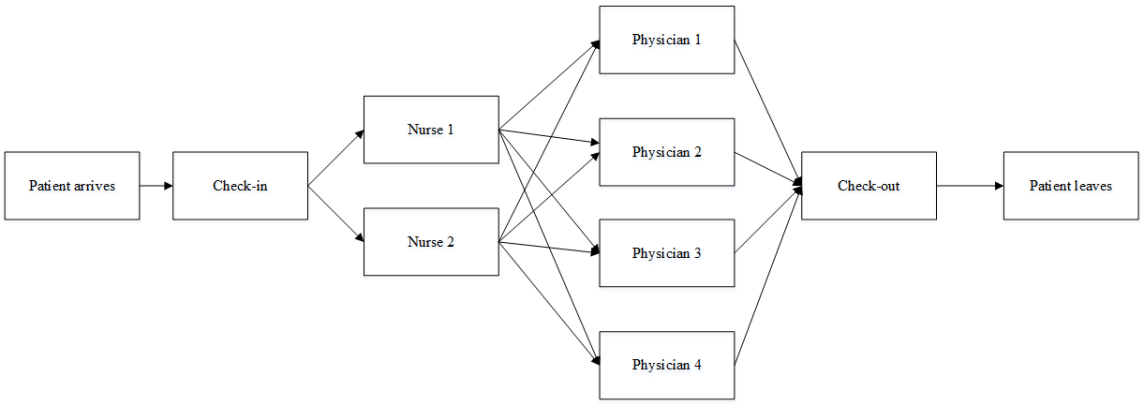


Figure 4.6: Illustration of multi-phase multi-provider system

Generalized Model Formulation

To facilitate better understanding of the mathematical model, we first formulate the model with the non-linear terms. Later, we present the linear transformation of all the non-linear constraints using the three linearization techniques discussed in Appendix A. The notations used to represent the parameters and variables involved in this problem are as follows:

Indices and Sets

$t, t' \in \mathscr{T}$: Set of patient types t or t'

$p \in \mathscr{P}_t$: Set of patients of patient type t

$s \in \mathscr{S}$: Set of slots

$h \in \mathscr{H}$: Set of phases

$r \in \mathscr{R}_h$: Set of resources in phase h

Parameters

$\alpha^t_{p,s}$: 1 if patient p of type t is available in slot s; 0 otherwise

$\rho^t_{p,r,h}$: 1 if patient p of type t prefers resource r in phase h ; 0 otherwise

η^t_p: Show rate of patient p of type t

$\nu^t_{p,h}$: Service time of patient p of type t in phase h

$EST_{h,s}$: Expected appointment start time of slot s for phase h

$ECT_{h,s}$: Expected appointment completion time of slot s for phase h

C^{IT}_h: Idle time cost for resources in phase h (\$/time unit)

C^{ST}_h: Spillover time cost for resources in phase h (\$/time unit)

C^{WT}: Patient waiting time cost (\$/time unit)

C^U: Cost of an unscheduled patient (\$/patient)

M: Very large positive number

Decision Variables

$X^t_{p,s}$: 1 if patient p of type t is scheduled to slot s; 0 otherwise

$Y^t_{p,r,h,s}$: 1 if patient p of type t who is scheduled to slot s is assigned to resource r in phase h; 0 otherwise

$B^t_{p,r,h}$: Start time of patient p of type t by resource r in phase h

$F^t_{p,r,h}$: Finish time of patient p of type t by resource r in phase h

$E_{r,h,s}$: Earliest time at which resource r in phase h can begin service in slot s

$L_{r,h,s}$: Latest time at which resource r in phase h completes the service in slot s

$AIT_{r,h,s}$: Idle time of resource r in phase h after completing service in slot s

$BIT_{r,h,s}$: Idle time of resource r in phase h before beginning service in slot s

$WT^t_{p,h}$: Waiting time of patient p of type t in phase h

$ST_{r,h,s}$: Spillover time of resource r in phase h of slot s

Mathematical Formulation

$$\text{Minimize } z = \begin{aligned} &\Big[\sum_{s\in\mathscr{S}}\sum_{h\in\mathscr{H}}\sum_{r\in\mathscr{R}_h} C^{IT}_h \times (BIT_{r,h,s}+AIT_{r,h,s})\Big] + \Big[\sum_{s\in\mathscr{S}}\sum_{h\in\mathscr{H}}\sum_{r\in\mathscr{R}_h} C^{ST}_h \times ST_{r,h,s}\Big] + \\ &C^{WT}\Big[\sum_{t\in\mathscr{T}}\sum_{p\in\mathscr{P}_t}\sum_{h\in\mathscr{H}} WT^t_{p,h}\Big] + C^U\Big[\sum_{t\in\mathscr{T}}\sum_{p\in\mathscr{P}_t}\sum_{s\in\mathscr{S}}(1-X^t_{p,s})\Big] \end{aligned} \tag{4.45}$$

s.t.

$$\sum_{s\in\mathscr{S}} X^t_{p,s} \le 1 \qquad t\in\mathscr{T},\ p\in\mathscr{P}_t \tag{4.46}$$

$$\sum_{r\in\mathscr{R}_h} Y^t_{p,r,h,s} = X^t_{p,s} \qquad t\in\mathscr{T},\ p\in\mathscr{P}_t,\ h\in\mathscr{H},\ s\in\mathscr{S} \tag{4.47}$$

$$Y^t_{p,r,h,s} \le \rho^t_{p,r,h} \times \alpha^t_{p,s} \qquad t\in\mathscr{T},\ p\in\mathscr{P}_t,\ h\in\mathscr{H},\ r\in\mathscr{R}_h,\ s\in\mathscr{S} \tag{4.48}$$

$$X^t_{p,s} + X^{t'}_{p',s} \le 1 \qquad t,t'\in\mathscr{T}, t\ne t',\ p\in\mathscr{P}_t,\ p'\in\mathscr{P}_{t'},\ s\in\mathscr{S} \tag{4.49}$$

$$\sum_{t\in\mathscr{T}}\sum_{p\in\mathscr{P}_t} Y^t_{p,r,h,s} \le 2 \qquad h\in\mathscr{H},\ r\in\mathscr{R}_h,\ s\in\mathscr{S} \tag{4.50}$$

$$\begin{aligned} B^t_{p,r,h} = \max\Big\{ &\Big(\sum_{s\in\mathscr{S}} EST_{h,s}\times Y^t_{p,r,h,s}\Big), \\ &\Big(\sum_{r'=1}^{R_{h-1}} F^t_{p,r',h-1} \times \sum_{s\in\mathscr{S}} Y^t_{p,r,h,s} : h>1,\ r'\in\mathscr{R}_h\Big), \\ &\Big(L_{r,h,s-1}\times X^t_{p,s} : s\in\mathscr{S}\ni s>1\Big), \\ &\Big(F^t_{p',r,h}\times Y^t_{p,r,h,s}\times X^t_{p',s} : p'\in\mathscr{P}_t\ni p'\le p-1, s\in\mathscr{S}\Big)\Big\}\ t\in\mathscr{T},\ p\in\mathscr{P}_t,\ h\in\mathscr{H},\ r\in\mathscr{R}_h \end{aligned} \tag{4.51}$$

$$B^t_{p,r,h} \leq M \sum_{s \in \mathscr{S}} Y^t_{p,r,h,s} \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h \tag{4.52}$$

$$F^t_{p,r,h} = (B^t_{p,r,h} + \eta^t_p \nu^t_{p,h}) \times \sum_{s \in \mathscr{S}} Y^t_{p,r,h,s} \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h \tag{4.53}$$

$$L_{r,h,s} = \max\left(F^t_{p,r,h} \times X^t_{p,s} : t \in \mathscr{T},\ p \in \mathscr{P}_t\right) \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \tag{4.54}$$

$$E_{r,h,s} = EST_{h,s} \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \ni s = 1 \tag{4.55}$$

$$E_{r,h,s} = \max\left(L_{r,h,s-1},\ EST_{h,s}\right) \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \ni s > 1 \tag{4.56}$$

$$BIT_{r,h,s} \geq L_{r,h,s} - E_{r,h,s} - \sum_{t \in \mathscr{T}} \sum_{p \in \mathscr{P}_t} \nu^t_{p,h} \times \eta^t_p \times Y^t_{p,r,h,s} \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \tag{4.57}$$

$$AIT_{r,h,s} - ST_{r,h,s} = ECT_{h,s} - L_{r,h,s} \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \tag{4.58}$$

$$WT^t_{p,h} = \sum_{r \in \mathscr{R}_h} B^t_{p,r,h} - \sum_{s \in \mathscr{S}} EST_{h,s} \times X^t_{p,s} \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H} \ni h = 1 \tag{4.59}$$

$$WT^t_{p,h} = \sum_{r \in \mathscr{R}_h} B^t_{p,r,h} - \sum_{r \in \mathscr{R}_h - 1} F^t_{p,r,h-1} \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H} \ni h > 1 \tag{4.60}$$

$$B^t_{p,r,h},\ F^t_{p,r,h},\ E_{r,h,s}, L_{r,h,s},\ BIT_{r,h,s}, AIT_{r,h,s},\ ST_{r,h,s},\ WT^t_{p,h} \geq 0 \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H}, r \in \mathscr{R}_h,\ s \in \mathscr{S} \tag{4.61}$$

$$X^t_{p,s},\ Y^t_{p,r,h,s} \in \{0,1\} \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \tag{4.62}$$

The objective function (Equation 4.45) seeks to minimize the weighted sum of resource idle time, resource spillover time, patient waiting time and denied appointment requests, where the weights are the respective costs.

Constraint (4.46) ensures that a patient is scheduled to at most one slot. Constraint (4.47) assigns the scheduled patient to exactly one resource at each phase. Constraint (4.48) ensures that a patient is scheduled only based on their physician preference and availability. Constraint (4.49) ensures that each slot can either accommodate a pre-book or an open access patient-type but not both. To avoid over-utilization of resources and increase in patient waiting time, Constraint (4.50) restricts the number of patients scheduled to a slot to no more than two.

It is important to ensure that the precedence relationship for each patient is satisfied. In other words, a patient can be served in phase h only after that patient completes his/her service in phase $h-1$. Moreover, a patient may not be immediately served upon arriving at a phase because the resource at that phase may be busy serving another patient. Therefore, if a patient is scheduled to slot s and assigned to resource r in phase h ($Y^t_{p,r,h,s} = 1$), then the service start time ($B^t_{p,r,h}$) will be equal to the maximum of the following four events:

(i) expected start time of phase h for slot s ($EST_{h,s}$) ,

(ii) service completion time of patient p in the previous phase $h-1$ ($\sum_{r'=1}^{R_{h-1}} F^t_{p,r',h-1}$). Note that $F^t_{p,r',h-1} = 0$ if resource r is not assigned to patient p.

(iii) latest service completion time of resource r in previous slot $s-1$ ($L_{r,h,s-1}$)

(iv) service completion time of resource r for patient p' ($F^t_{p',r,h}$), who is scheduled before patient p in the same slot ($X^t_{p',s} = 1$) and assigned to resource r ($Y^t_{p',r,h}$=1).

The events (i) - (iv) are active only if all the binary variables associated with that event is equal to 1. In other words, events (i) and (ii) are active only if $\sum_{s\in\mathscr{S}} Y^t_{p,r,h,s} = 1$, event (iii) is active only if $X^t_{p,s} = 1$, and event (iv) is active only if $Y^t_{p,r,h,s} = X^t_{p',s} = 1$. Thus, the begin time, $B^t_{p,r,h}$ is a non-linear term as shown in Equation (4.4) and can be linearized using Constraints (4.63) - (4.66).

$$B^t_{p,r,h} \geq \sum_{s\in\mathscr{S}} EST_{h,s} \times Y^t_{p,r,h,s} \qquad t\in\mathscr{T},\ p\in\mathscr{P}_t,\ h\in\mathscr{H},\ r\in\mathscr{R}_h \tag{4.63}$$

$$B^t_{p,r,h} \geq \sum_{r'=1}^{R_{h-1}} F^t_{p,r',h-1} - M(1-\sum_{s\in\mathscr{S}} Y^t_{p,r,h,s}) \qquad t\in\mathscr{T},\ p\in\mathscr{P}_t,\ h\in\mathscr{H}\ni h>1,\ r\in\mathscr{R}_h \tag{4.64}$$

$$B^t_{p,r,h} \geq L_{r,h,s-1} - M(1-X^t_{p,s}) \qquad t\in\mathscr{T},\ p\in\mathscr{P}_t,\ h\in\mathscr{H},\ r\in\mathscr{R}_h,\ s\in\mathscr{S}\ni s>1 \tag{4.65}$$

$$B^t_{p,r,h} \geq F^t_{p',r,h} - M(2-Y^t_{p,r,h,s}-X^t_{p',s}) \qquad t\in\mathscr{T},\ p,p'\in\mathscr{P}_t,\ p'\leq p-1,\ h\in\mathscr{H},\ r\in\mathscr{R}_h,\ s\in\mathscr{S} \tag{4.66}$$

However, if patient p is not assigned to resource r in phase h (i.e., $\sum_{s\in\mathscr{S}} Y^t_{p,r,h,s} = 0$), then Constraint (4.52) forces the begin time of patient p by that resource to zero.

At phase h, if patient p is assigned to resource r ($\sum_{s\in\mathscr{S}} Y^t_{p,r,h,s} = 1$), then the service completion time of that patient in that phase is the sum of service start time and the expected service time. However, if the patient is not assigned to a resource in that phase ($\sum_{s\in\mathscr{S}} Y^t_{p,r,h,s} = 0$), then the completion time is zero. This condition leads to constraint with a non-linear term as shown in Equation (4.53). We use a linearization technique to replace Constraint (4.53) with Constraints (4.67) - (4.69).

$$F^t_{p,r,h} \leq B^t_{p,r,h} + \eta^t_p v^t_{p,h} + M(1 - \sum_{s\in\mathscr{S}} Y^t_{p,r,h,s}) \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h \tag{4.67}$$

$$F^t_{p,r,h} \geq B^t_{p,r,h} + \eta^t_p v^t_{p,h} - M(1 - \sum_{s\in\mathscr{S}} Y^t_{p,r,h,s}) \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h \tag{4.68}$$

$$F^t_{p,r,h} \leq M \sum_{s\in\mathscr{S}} Y^t_{p,r,h,s} \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h \tag{4.69}$$

A slot can be single booked or double-booked, and the resources must serve all the patients assigned to them in a slot before beginning service in the next slot. Further, the total expected service time of the patient(s) scheduled in a slot may be longer or shorter than the slot duration. These factors impact the latest time at which a resource completes the service and the earliest time at which a resource can begin service in a slot. The latest resource completion time for a slot is the completion time of the last patient scheduled to that slot, and is given by Constraint (4.54). The non-linearity in Constraint (4.54) can be avoided by replacing it with Constraint (4.70).

$$L_{r,h,s} \geq F^t_{p,r,h} - M(1 - X^t_{p,s}) \qquad t \in \mathscr{T},\ p \in \mathscr{P}_t,\ h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \tag{4.70}$$

The earliest resource start time for the first slot is the expected appointment start time of the first slot, and is given by Constraint (4.55). The earliest resource start time for all the other slots will be the maximum of two times, namely, latest completion time of the resource in the previous slot and the expected start time of the current slot, and is given by Constraints (4.56). This non-linear constraint is transformed to linear Constraints (4.71) -

(4.74) by introducing a binary variable ($\delta_{r,h,s}$).

$$E_{r,h,s} \geq EST_{h,s} \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \ni s > 1 \tag{4.71}$$

$$E_{r,h,s} \geq L_{r,h,s-1} \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \ni s > 1 \tag{4.72}$$

$$E_{r,h,s} \leq EST_{h,s} + M(1-\delta_{r,h,s}) \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \ni s > 1 \tag{4.73}$$

$$E_{r,h,s} \leq L_{r,h,s-1} + M\delta_{r,h,s} \qquad h \in \mathscr{H},\ r \in \mathscr{R}_h,\ s \in \mathscr{S} \ni s > 1 \tag{4.74}$$

At a given slot, the resource can be idle in two situations: (i) before beginning the service and (ii) after completing the service. If the resource waits for the patient to arrive (e.g., waiting for service completion of the patient in the previous stage), then the resource is idle before beginning the service, and is given by Constraint (4.57). On the other hand, if the latest completion time of a resource in a phase is earlier than the expected completion time of that phase, then the resource is idle after completing the service. However, if the latest service completion time of a resource in a phase exceeds the expected completion time of that phase, then the resource is overburdened in that slot. The work overload in a slot is measured using the spillover time (additional time beyond the expected duration, which is required to complete the treatment of the patient(s) scheduled in that phase). The resource idle time after service completion and resource spillover time for each slot is given by Constraint (4.58).

The waiting time of patient for a resource in the first phase (Equation (4.59)) is evaluated as the difference between the service start time and the appointment start time for that phase. The patient waiting time for all other phases (Equation 4.60) is difference between the service start time in that phase and service completion time in the previous phase. The non-negativity and binary restrictions on the decision variables are ensured using Constraint (4.61) and Constraint (4.62), respectively. Therefore, in the deterministic MILP model, the objective function (Equation 4.45) will be subject to Constraints (4.46) - (4.50), (4.52), (4.55), and (4.57) - (4.74).

The generalized model is an extension of the two-phase model and has similar constraints. The scenario generation, heuristic procedure, and the solution approach for a multi-phase system is similar to the procedure of the two-phase system discussed in Sections 4.1 and 4.2. However, the data generated for each scenario may vary depending on the number of phases. For example, if there are three phases, then the service time for all the three phases must be generated.

4.5 Conclusions

In practice, most outpatient departments have a multi-phase multi-provider setting. Each patient is available to visit the clinic only during certain times of the day and has his(her) own provider preference. A patient's continuity of care is improved if he(she) is scheduled to see his(her) preferred physician. On the other hand, nurses and physicians are often overloaded with patients resulting in medical errors, fatigue, burnout and emotional stress. In addition, the uncertainties associated with the schedule, such as, patient no-show, patient service time, patient calls for open access and pre-book appointments lead to ineffective resource utilization and impacts patient satisfaction. However, most of the existing research focuses on determining the configuration of a single-phase single-provider system and do not consider factors, such as, patient availability, patient's provider preference and other uncertainties.

The proposed methodology in this chapter aims to address the aforementioned gaps in the literature and aids the hospital administrators in designing a hybrid appointment system. The research methodology uses a combination of mathematical modeling, scenario analysis and heuristics. First, an MILP model for a multi-phase multi-provider system is developed to determine the optimal configuration of the hybrid appointment system for a given scenario. The uncertainties associated with the outpatient department is handled by considering multiple scenarios, where each scenario serves as an input to the mathematical model and the optimal schedule changes depending on the scenario. The final configuration of the hybrid appointment system is determined using two heuristic approaches, namely,

Frequency method and Averaging method. The performance of the two heuristic approaches is determined using backtesting, where a simulation model is developed to evaluate the configuration obtained by each heuristic.

A case study is used to show the feasibility of this model and the heuristic approach. Real data from a Family Medicine Clinic located in Pennsylvania, USA is used for the case study. The model is solved with CPLEX and 20 different scenarios are analyzed to determine the final configuration of the hybrid appointment system for all the physicians. It is observed that the configuration obtained using Heuristic 1 (Frequency method) performs better with respect to the average resource overtime, patient waiting time, and overall cost, while Heuristic 2 (Averaging method) performs well with respect to average resource idle time and number of unscheduled patients.

Chapter 5

Data-Driven Appointment Rules for Outpatient Scheduling using Patient Specific No-Show Rates

5.1 Introduction

The mathematical model discussed in Chapter 4 focuses on determining the schedule configuration (i.e., percentage of appointments reserved for each appointment type and their position in the schedule) of a hybrid appointment system. Once the configuration is established, it will be in use as long as the system parameters (e.g., patient demand, no-show rate) doesn't change drastically. However, it is important to establish policies or rules to schedule individual patients as and when they call for appointment. In this chapter, different data-driven scheduling rules are proposed and evaluated to schedule patients efficiently when they call for appointments.

As discussed in Chapter 1, the demand for outpatient visits are rising and hospitals are expected to receive 65% of their revenue from outpatient care. The increase in demand is primarily due to the following reasons:

- Introduction of the affordable care act resulting in 33 million Americans with access to health insurance
- Shift from the inpatient service to outpatient services as a result of new technologies, reimbursement rules and payment models.

- Aging population that requires constant care

On the other hand, hospitals are expected to have a shortage of 90,000 physicians by 2025. Thus, an increase in demand for services, coupled with shortages in the supply of physicians, are leading to a supply-demand mismatch. The supply demand mismatch would have negative impact on both patient satisfaction and hospital revenue. It would result in increased patient waiting times and long appointment lead times (i.e., the time between the call for appointment and actual appointment date). Fearing high waiting times, the patient may not show for the appointment, which would lead to inefficient resource utilization, thereby decreasing the hospital revenue.

To improve the resource utilization, patient satisfaction and revenue generated, outpatient clinics generally use an appointment system to distribute the workload (demand) throughout the day by scheduling patient to smaller time duration called slots. However, several environmental factors (uncertainties) affect the appointment system leading to decrease in patient satisfaction and poor utilization of resources. One of main causes of uncertainty is patient no-shows, where a patient misses an appointment without prior notice. The average no-show rates for primary care clinics vary between 14% and 50% (Daggy et al., 2010). It is estimated that patient no-shows cost healthcare system $150 billion a year (Toland, 2013).

To handle the increase in demand and decrease in revenue as a result of patient no-shows, hospitals tend to overbook. In other words, they schedule more than their capacity to compensate for patient no-shows. Most hospitals overbook by using a flat overbooking percentage which often depends on the average no-show rate of the clinic. For example, consider a clinic that has a capacity to serve 20 patients per day (i.e., 20 slots) and an average no-show rate of 20%. If the clinic uses the average no-show rate to determine the flat overbooking value, then the clinic will schedule 25 patients per day ($\frac{20}{1-0.2}$) resulting in a flat overbooking value of 5 patients (25 patients scheduled - 20 patient capacity = 5 overbooked patient). Therefore, if the no-show rate is 20% for a given day, then the number of patients arriving for that day will be exactly equal to the clinic's daily capacity (i.e, 25 x

(1-0.2) = 20).

Prior research have studied the impact of using an overbooking approach on no-shows. LaGanga and Lawrence (2007) are the first to focus on the use of overbooking as a technique to minimize the negative consequence of no-shows. Based on their simulation study, it was found that patient access and provider productivity significantly improved with overbooking at the expense of patient waiting time and provider overtime. However, their approach considered all the patients to have the same no-show probability.

Some research also considered heterogeneous patients (i.e., patient with different no-show probabilities). Muthuraman and Lawley (2008) proposed a stochastic overbooking model and a myopic scheduling policy by categorizing patients into groups based on their estimated no-show probability and scheduling patients in a sequential fashion as and when the call arrives. Zeng et al. (2010) proposed an overbooking policy for a set of heterogeneous patients with the objective of maximizing expected profit, and provided managerial insights based on their analysis. Huang and Zuniga (2012) proposed a dynamic overbooking policy by considering the probability of patient no-shows. The authors concluded that the total cost of patient waiting time, physician idle time and over time can be reduced when overbooking is practiced for high no-show rates and when the appointment duration is longer than the mean service times. Recently, Samorani and LaGanga (2015) examined the combined use of analytics, optimization and overbooking to schedule appointments in the presence of no-shows. One of their major findings is that the prediction quality of the machine learning algorithm should be shifted towards high specificity (probability of predicting show patients (low-risk) correctly) for clinics adopting heavy overbooking and towards high sensitivity (probability of predicting no-show patients (high-risk) correctly) for clinics adopting light overbooking.

Some studies proposed scheduling policies by using patient specific no-shows. Glowacka et al. (2009) predicted patient no-shows, using association rule mining and then scheduled patients in ascending order of their no-show probabilities. However, their model assigned no-show values only to some patient groups and did not consider a sequential call-in

process. Daggy et al. (2010) predicted patient-specific no-show values using a logistic regression model and scheduled patients based on their no-show probabilities to balance patient waiting time, clinic over time and revenue. They compared their model with the practice of scheduling one patient in each slot without regard to their no-show probability and concluded that scheduling models that considered individual patient no-show values might improve the schedule efficiency, without limiting access to patient care.

The current scheduling practice at most hospitals and some of the models proposed in the literature assume the patients to be homogeneous (i.e., have the same no-show probability). In addition, the current practice randomly overbooks patients to predetermined slots without considering the risk of collisions (i.e., both patients showing up for an appointment). However, each patient may have a specific no-show value depending on various attributes, such as age, race, appointment time, appointment length, and distance to clinic. Further, research that considers the individual no-show probability, does not consider the actual nature of patient flow, multi-provider clinic setting and variable appointment durations. Based on the interaction with multiple healthcare facilities, it is observed that outpatient clinics employ multiple providers and generally have a multi-phase setting, where the patient moves through a sequence of phases (e.g., check-in, visit nurse, visit physician, check-out) before exiting the system. As discussed in Chapter 3, ignoring the nature of patient flow will alter the values of performance measures, such as patient waiting time and resource idle time. The objective of this chapter is to overcome these gaps in the literature and address the issues with the current practice.

This chapter contributes to the literature in four ways. First, we develop a framework to improve the performance of an appointment system with respect to patient satisfaction and resource utilization. The proposed framework aims to exploit the available data and obtain meaningful insights about the patient no-shows, and then uses it to design an easy-to-implement appointment system. Even though prior research have used existing data to estimate no-show probabilities and have used it to design a scheduling policy, the same methods/approach may not be suitable for all clinics. The generalized framework proposed

in this chapter enables the clinics to adapt it to their clinic setting. Second, data from multiple sources and unique features, such as weather conditions, are used as inputs to predict no-shows. To the best of our knowledge, none of the research work has used weather conditions to estimate no-show probability. The advancement in science and technology have paved way to obtain reasonably accurate weather forecasts, and we have used it to our advantage in predicting no-shows. Third, the findings in the literature have been adapted to develop new sequential scheduling rules, where patients are scheduled as and when they call for appointment. Finally, we consider realistic clinic settings, such as multi-phase multi-provider clinics and variable appointment duration, to evaluate the proposed scheduling rules.

5.2 Methodology

The proposed appointment scheduling framework is illustrated in Figure 5.1. The patients are assumed to be non-homogeneous, with different no-show rates. The first step is to obtain patient-related data from various sources, such as electronic health records (EHR). The next step is to process the data and use supervised machine learning (ML) algorithms to uncover patterns in the underlying data and classify the patients based on their no-show rates as high-risk or low-risk. A high-risk patient is very likely to miss the appointment, while a low-risk patient is very likely to arrive for the scheduled appointment. In the last step, new scheduling rules that leverages the patient information from predictive analytics are proposed to schedule patients as they call in for appointments.

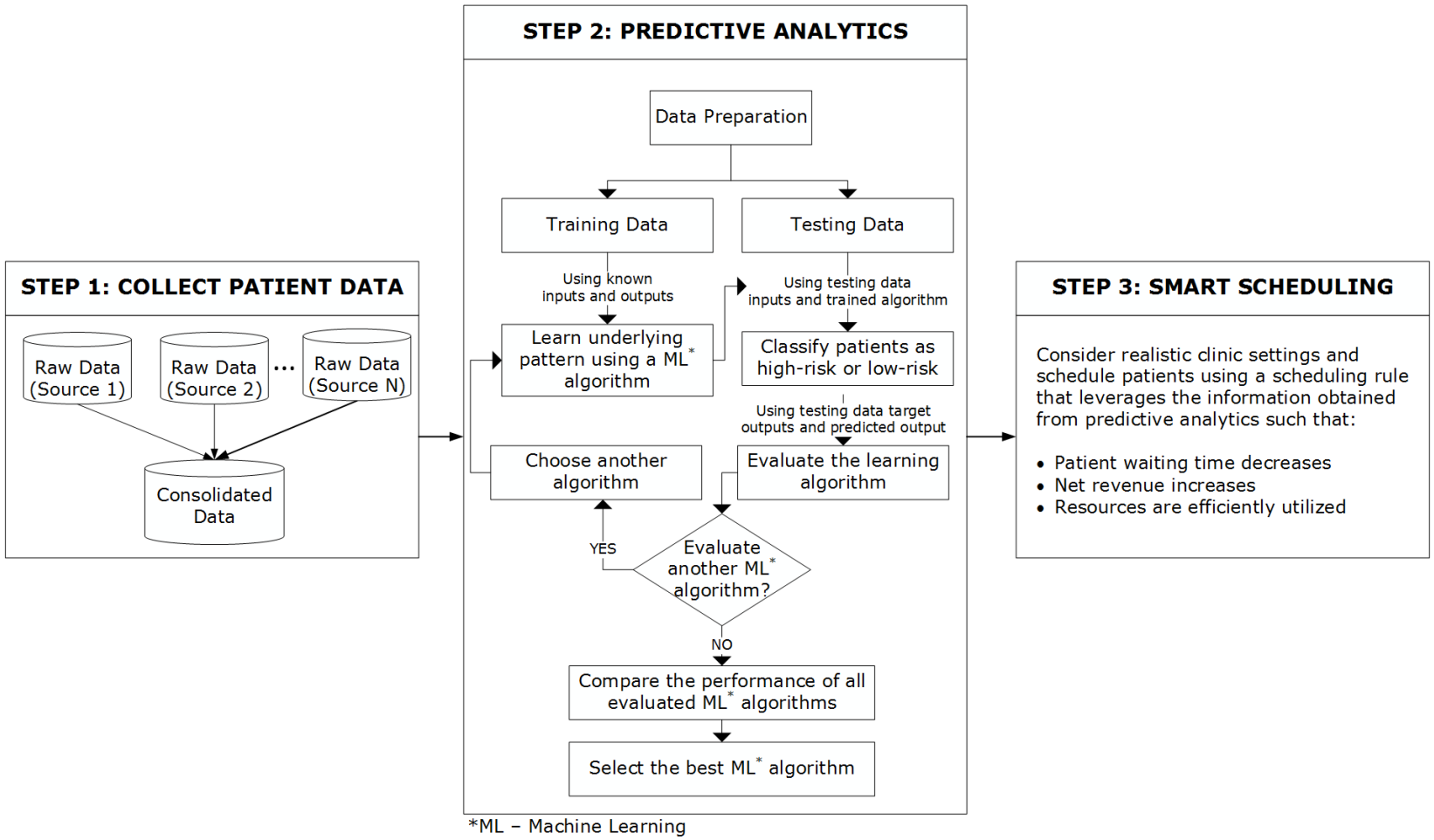


Figure 5.1: Proposed framework for outpatient scheduling

5.2.1 Data Collection

We illustrate the proposed framework with actual patient dataset. The historical data was obtained from the scheduling system of a Family Medicine Clinic in Central Pennsylvania and included the electronic health records (EHR) of patient visits from September 2014 to August 2016. The entire dataset included 76,285 patient visits. The data elements are chosen based on their availabilities and with consultation from the clinical transformation team at the clinic. In addition to EHR data, certain other data were obtained from public data sources. For example, using patient's zip code and the clinic's zip code information, the driving distance to clinic was calculated using Google Maps API. Similarly, using the patient's zip code, the weather conditions, such as temperature and precipitation, are extracted from a website providing real-time and historical weather information. Thus, data from multiple sources were obtained and consolidated.

The consolidated dataset has 18 features (independent variables) and can be grouped into three main categories:

1. **Patient Information**: Provides information about the patient visiting the clinic such as age, race, gender, and insurance provider.

2. **Appointment Information**: Provides information relating to the appointment such as time, duration and month of the appointment.

3. **Weather Information**: Provides information related to the weather such as temperature and chances of rain.

The response variable (dependent variable) is the patient no-show value. For the purpose of this study, patients are considered as no-show, if they missed their appointments without notice or canceled their appointments within 72 hours prior to their visit. A detailed list of all the features in the dataset is given in Table 5.1. The variables Max Temperature, Min Temperature, Lead Time, Precipitation Probability and Precipitation Intensity are treated as continuous variables, while the remaining variables are set as categorical variables

(i.e., a variable that can take on a limited number of pre-specified values). For example, the variable “Duration" is a categorical variable, as it can only take one of the three values - Brief, Intermediate and Extended. On the other hand, the variable “Lead Time" is continuous because it can take any value and is not fixed.

Each categorical independent variable is coded by creating dummy variables which use 0 and 1 to represent the category information. Further, to avoid the issue of multicollinearity in methods such as logistic regression, $k-1$ dummy variables are created for a variable with k categories. For example, the variable “Age" has 5 categories as shown in Table 5.1, and requires 4 dummy variables to represent all the categories. For example, if the age of the patient is between 0 and 20 years, then $A_1 = 1$, $A_2 = 0$, $A_3 = 0$, and $A_4 = 0$. Similarly, if the age of the patient is between 81 and 100 years, then all A_i's are zero (i.e., $A_1 = 0$, $A_2 = 0$, $A_3 = 0$, and $A_4 = 0$). Thus, only $k - 1$ dummy variables are necessary to represent all the k categories of a categorical variable. Therefore, the categories of all the 13 categorical variables can be represented using 43 dummy variables as shown in Table 5.2. In addition, there are 5 continuous variables. Thus, there are 48 variables (43 dummy + 5 continuous) in total.

Table 5.1: Description of variables

Category	**Variable**	**Description**
Patient Information	Age (0-20, 21-40, 41-60, 61-80, 81-100)	Age of the patient
	Sex (Male, Female)	Gender of the patient
	Marital Status (Single, Divorced, Married, Separated, Widowed)	Marital status of the patient
	Race (African American, Asian, Caucasian, Other)	Race of the patient
	Driving Time ($<=$ 10, 11-20, 21-30, $>$ 30)	Driving time to the clinic (in minutes)
	Insurance Group (Medicare, Medicaid, Private, Uninsured)	Insurance group of the patient
	Patient Type (New, Return)	Patient category based on the visit number
	Visit Type (Lab, Acute, Other)	Purpose of the visit to the clinic
Appointment Information	Month (Jan, Feb, Mar,...,Dec)	Month in which appointment is scheduled
	Day (Mon, Tue, Wed, Thu, Fri)	Day on which the appointment is scheduled
	Session (Morning, Forenoon, Afternoon, Evening)	Session in which the appointment is scheduled
	Duration (Brief, Intermediate, Extended)	Duration of the appointment
	Lead Time	Days between patient call for appointment and appointment date
Weather Information	Max Temperature	Forecasted maximum temperature (in °F)
	Min Temperature	Forecasted minimum temperature (in °F)
	Precipitation Probability	Chances of precipitation
	Precipitation Intensity	Intensity of precipitation (inches/hour)
	Precipitation Type (Snow, Rain)	Type of precipitation

Table 5.2: Coding of categorical variables

Categorical Variable	**Dummy Variable**	**Description**
Age	A_x	Age of patient is between x years, where x = 1,2,3,4 (1 - 0 to 20, 2 - 21 to 40, 3 - 41 to 60, 4 - 61 to 80)
Sex	G	Gender of patient is male (1 - Male, 0 - Female)
Marital Status	M_i	Marital status of patient is i,where i = 1,2,3,4 (1 - Single, 2 - Divorced, 3 - Married, 4 - Separated)
Race	R_r	Patient belongs to race r, r = 1,2,3,4 (1 - African American, 2 - Asian, 3 - Caucasian)
Driving Time	T_t	Driving time is t minutes, t = 1,2,3 (1 - <= 10, 2 - 11 to 20, 3 - 21 to 30)
Insurance	I_g	Patient has g insurance, where g = 1,2,3 (1 - Medicare, 2 - Medicaid, 3 - Private)
Patient Type	P	Patient visiting for the first time (1 - New patient, 0 - Returning Patient)
Visit Type	V_v	Patient scheduled for v visit, where v = 1,2 (1 - Lab, 2 - Acute care)
Month	M_m	Patient scheduled for month m, where m = 1,2,...,11 (1 = Jan, 2 = Feb,...)
Day	D_d	Patient scheduled on day d, where d = 1,2,...,4 (1 = Mon, 2 = Tue,...)
Session	S_s	Patient scheduled for session s, where s = 1,2,3 (1 - Morning, 2 - Forenoon, 3 - Afternoon)
Duration	L_l	Appointment duration of length l, where l = 1,2 (1 - Short, 2 - Medium)
Precipitation Type	P	Precipitation type on day of appointment (1 -Rain , 0 - Snow)

5.2.2 Classifying Patients using Predictive Analytics

Several machine learning (ML) algorithms are used to classify the patients based on their risk type, as either high-risk (high probability of being a no-show) or low-risk. The entire dataset is divided into training and testing dataset. The ML learning algorithm learns the underlying relationships between the independent variables (Table 5.1) and the dependent variable (high or low risk) using the training dataset, where the inputs (values of independent variables) and the expected output (value of dependent variable) are presented to the ML algorithm for learning. Further, to validate the model and avoid the risk of overfitting (i.e., model learning the noise in the data), techniques such as k-fold cross validation is used in the training phase. In k-fold cross validation, the training dataset is split into k-subsets, and one of the subset is set aside as validation data and the remaining k-1 subsets are used for training the data. This process is repeated k times, where each of the k subsets are used exactly once as validation data.

Later, the trained ML algorithm evaluates the underlying relationship by using only the inputs from the testing dataset to predict the dependent variable (output). Finally, the performance of the ML algorithm is evaluated by comparing the predicted output and the actual output.

Five different machine learning algorithms have been used in this section and the best method for classifying the patients based on their risk type is identified. The five methods considered are: (i) Logistic Regression (ii) Neural Networks (iii) Random Forests (iv) Gradient Boosting (v) Stacking. Brief descriptions of the ML algorithms are given in the next sub-sections.

5.2.2.1 Logistic Regression

The probability that the patient will be of high-risk or low-risk is computed using the independent variables and is shown in Equations 5.1 and 5.2, where b_0 is a constant and b_i's are the regression coefficients of the input parameters. The training dataset is used to estimate the regression constant and the regression coefficients that best fit the observed data.

$$\text{P(Risk Type = High)} = \frac{1}{1 + e^{-(b_0 + b_1 * A_1 + b_2 * A_2 + ... + b_{48} * P)}} \tag{5.1}$$

$$\text{P(Risk Type = Low)} = 1 - \text{P (Risk Type = High)} \tag{5.2}$$

A schematic representation of the relationship between the input features and the response variable with a logistic function is shown in Figure 5.2.

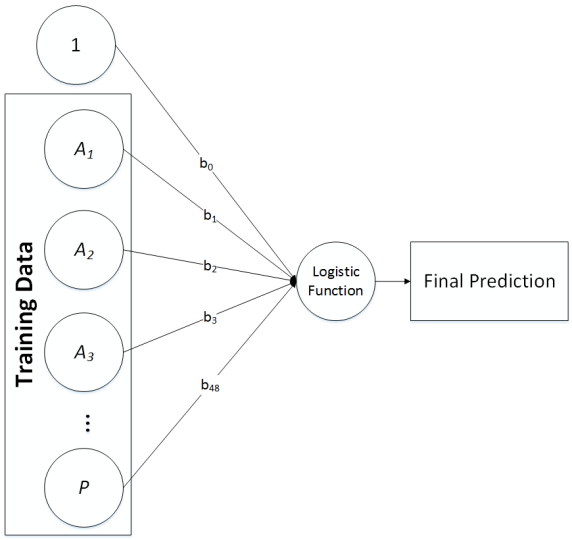


Figure 5.2: Schematic representation of Logistic Regression

5.2.2.2 Artificial Neural Network (ANN)

Artificial neural network (ANN) is a machine learning algorithm that is inspired by the biological neural network. As shown in Figure 5.3, ANN includes three different types of layers, namely, an input layer, a hidden layer, and an output layer and each layer has a certain number of nodes. Each node (i) in a given layer (l) is connected to each node (j) in the next layer (l+1) by a connection weight (w_{ij}). In order to train the classifier, the patient features are given as inputs to the input layer. Each input value is multiplied by a weight at the input layer, and the weighted input is relayed to each node in the hidden layer. Each node in the hidden layer will combine the weighted inputs that it receives, use it with the activation function (e.g., sigmoid activation), and relay the value to the nodes in the output layer. The output layer then determines the network output (risk type of

patient) by performing a weighted sum of the outputs of the hidden layer. The process of using the training inputs, hidden layers and activation function to compute the dependent variable (risk type) is called feed-forwarding or forward-pass. Initially, the training process begins with random weights. At the end of each feed-forward step, the predicted output is compared with the actual output. If the predicted risk type is same as the actual risk type, then the neural network's weights are reinforced. On the other hand, if the predicted risk type is incorrect, then the neural network's weights are adjusted based on a feedback. This process is called backward-pass. The forward-pass and the backward-pass are repeated for different training samples until the ANN classifier is fully trained. ANN can be useful to uncover complex relationship between the inputs and the output. However, it requires more parameters to be estimated, and therefore, may require more time for training.

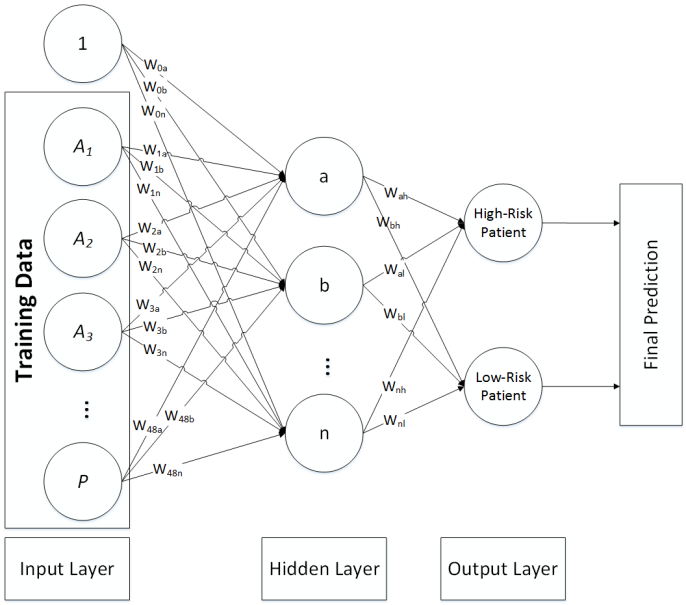


Figure 5.3: Schematic representation of Artificial Neural Networks

5.2.2.3 Random Forests (RF)

Random forests is an ensemble of decision trees proposed by Breiman (2001). Decision trees use a tree-like structure to obtain the output class and have nodes at each level of the tree. Each node splits into two or more nodes in the next level and the dataset is divided

among the nodes based on a test (e.g., is variable 'temperature' $>$ 10 ?). This process is repeated until the output class (risk-type of patient) is reached. At each level, it is necessary to select the independent variable that is most useful for classifying the dependent variable, and the information gain is a metric to measure the usefulness of a patient feature at that level. Information gain is the expected reduction in entropy due to sorting on a given node. Entropy is the measure of impurity and a higher entropy indicates more information content. If p_i indicates the probability of class i, then Equations 5.3 and 5.4 gives the entropy and information gain respectively.

$$\text{Entropy} = \sum_i -p_i log_2 p_i \tag{5.3}$$

$$\text{Information gain = Entropy(parent)} - \text{Weighted Average [Entropy(children)]} \tag{5.4}$$

In random forests, multiple decision trees are simultaneously trained, where each decision tree provides an output class. The final output class is the plurality voting of all the decision trees. Figure 5.4 is a schematic representation of the random forest algorithm. Decision trees do not use a linear combination of the independent variables to predict the output. Therefore, the independent variables need not be dummy coded when using decision trees. In fact, it is better to use the categorical variable as a single feature so that the decision tree can decide the factor levels on each side of the split. However, it is necessary to clearly declare the categorical and continuous variables before training the random forest algorithm, so that it can differentiate the different variable types.

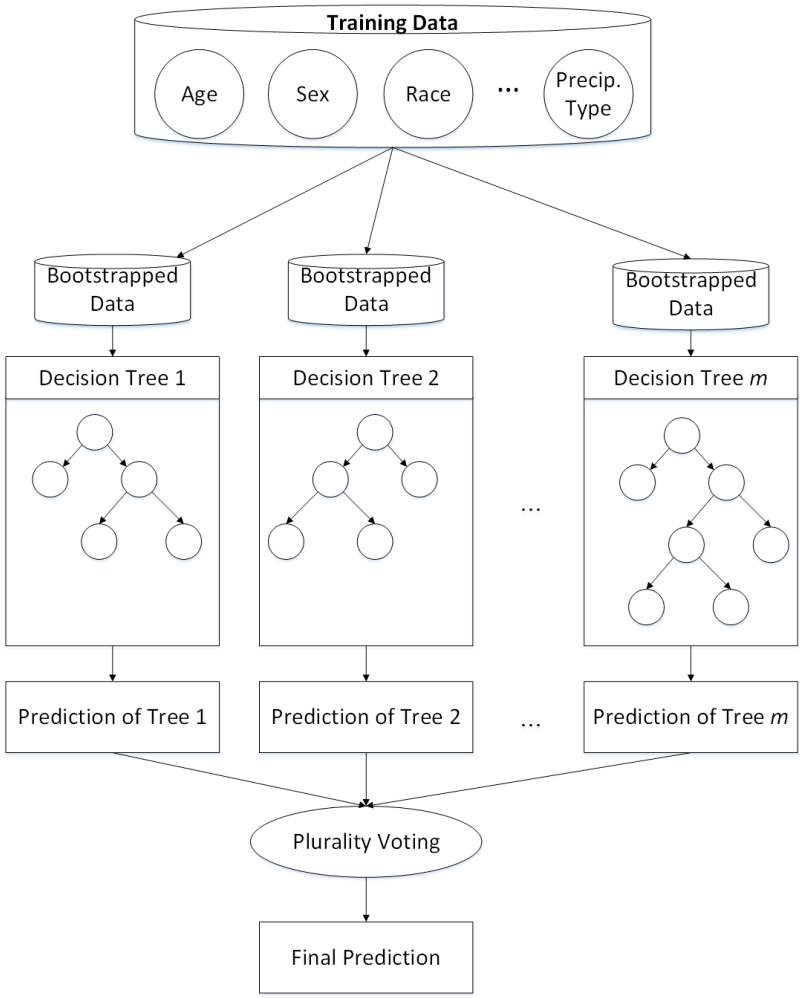


Figure 5.4: Schematic representation of Random Forests

5.2.2.4 Stochastic Gradient Boosted Decision Trees (SGBDT)

Stochastic gradient boosting is an ensemble method introduced by Friedman (2002). It iteratively trains shallow decision trees with the objective of minimizing a loss function (e.g., negative log-likelihood) and sequentially learns from the errors of the previous trees. Therefore, the trees are trained one at a time and cannot be trained in parallel. A schematic representation of SGBDT algorithm is shown in Figure 5.5.

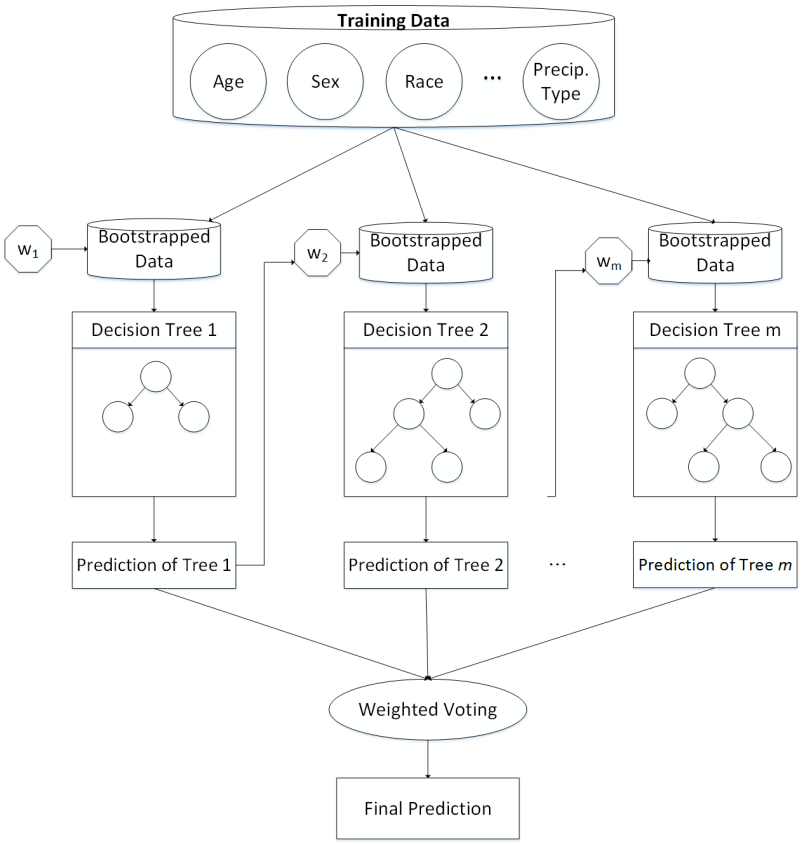


Figure 5.5: Schematic representation of SGBDT

The samples that are wrongly classified by a decision tree are up-weighted, and the samples that are correctly classified are down-weighted. This procedure is continued iteratively for all the decision trees leading to a higher weight for observations with correctly classified outputs and a lower weight for observations with wrongly classified outputs. It is to be noted that the SGBDT randomly samples a subset of the training data to train each decision tree. The final output would be a weighted voting of the decision trees.

5.2.2.5 Stacking

Stacking is also an ensemble method, where the predictions of multiple machine learning (ML) algorithms are combined. Stacking involves three stages, and the training dataset is split into two disjoint sets (training data-1 and training data-2). In the first stage, the training data-1 is used to train different ML algorithms (e.g., logistic regression, random forests),

where the independent variables serve as inputs. The ML algorithms used to train the independent variables are called base-level classifiers. In the second stage, training data-2 is used to obtain predictions (i.e., testing of base-level classifiers) from the trained base-level classifiers. In the third stage, the predicted outputs obtained from the base-level classifiers are used as inputs to a ML algorithm (e.g., ANN, random forests). The ML algorithm that uses the predictions of base-level classifiers as inputs is called meta-level classifier. Thus, the third stage combines the individual predictions of the base-level classifiers using a meta-level classifier to obtain the final class output. A schematic representation of the stacking algorithm is shown in Figure 5.6.

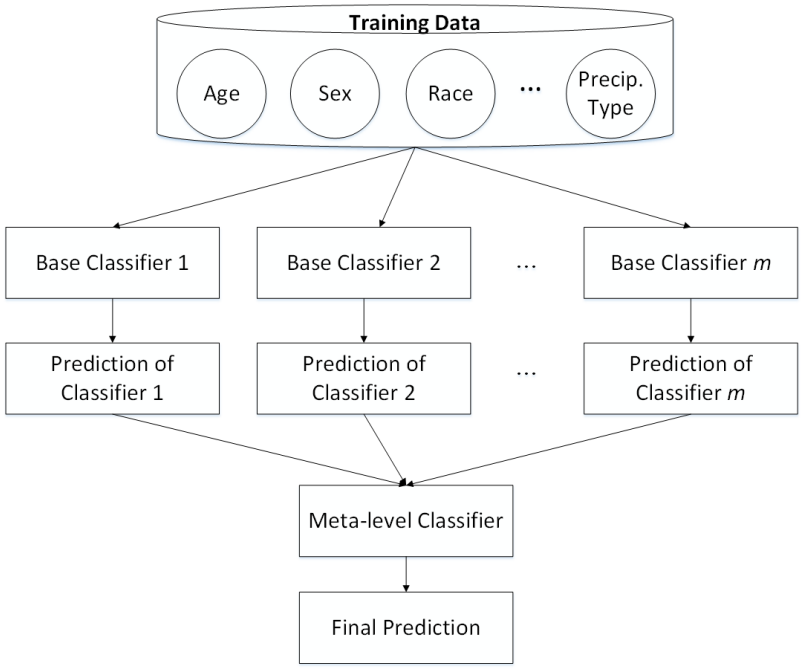


Figure 5.6: Schematic representation of Stacking

5.2.3 Evaluating a Machine Learning Algorithm

Given the set of input values, the trained ML algorithm provides the probability value for the patient risk type. A threshold parameter is then used to convert the probability to high-risk or low-risk patient. A patient is classified as low-risk if the probability is less than the threshold parameter and high-risk otherwise. The accuracy of the ML algorithm is

determined by using the actual output and the predicted output by constructing a confusion matrix. A model has high accuracy if the actual outputs are the same as the predicted outputs for many instances. As shown in Table 5.3, the confusion matrix has four categories, namely, true positive, true negative, false positive and false negative. If a ML algorithm predicts the risk type of a patient as high risk and if the actual risk type of the patient is low risk, then it is a false positive or type I error. Likewise, if a ML algorithm predicts the risk type of a patient as low risk and if the actual risk type of the patient is high risk, then it is a false negative or type II error.

Table 5.3: Confusion matrix

		Actual Output	
		High Risk	**Low Risk**
Predicted Output	**High Risk**	True Positive	False Positive (Type I error)
	Low Risk	False Negative (Type II error)	True Negative

True positive and true negative represent correct predictions by the ML algorithm. The true positive rate (TPR) is the percentage of high-risk patients that are correctly classified, while the true negative rate (TNR) is the percentage of low-risk patients that are correctly classified. Therefore, three metrics, namely, true positive rate (sensitivity), true negative rate (specificity) and accuracy, are derived from the confusion matrix and are given by Equations 5.5 - 5.7.

$$\text{True Positive Rate (Sensitivity)} = \frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Negative}} \tag{5.5}$$

$$\text{True Negative Rate (Specificity)} = \frac{\sum \text{True Negative}}{\sum \text{True Negative} + \sum \text{False Positive}} \tag{5.6}$$

$$\text{Accuracy} = \frac{\sum \text{True Positive} + \sum \text{True Negative}}{\sum \text{True Positive} + \sum \text{True Negative} + \sum \text{False Positive} + \sum \text{False Negative}} \tag{5.7}$$

The Receiver Operating Characteristics (ROC) curve plots the true positive rate (TPR) versus the false positive rate (1 - TNR) as the threshold parameter for classifying high-risk or low-risk patient is varied from 0 to 1 as shown in Figure 5.7. Therefore, each point on

the ROC curve represents the TPR/(1-TNR) value corresponding to a particular threshold.

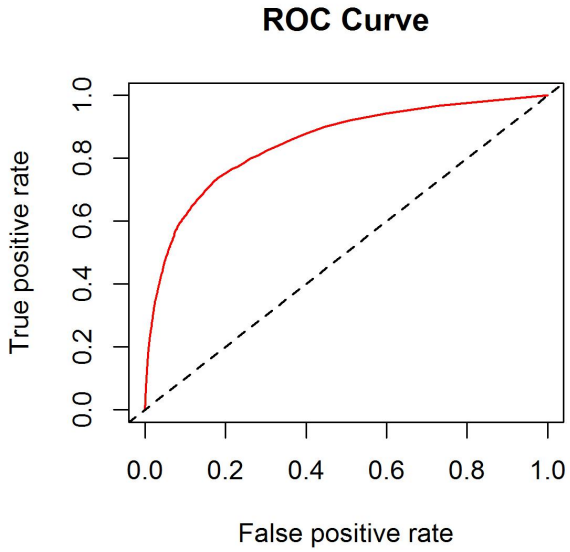


Figure 5.7: A sample ROC Curve

The area under the ROC curve, called AUC, is a single-measure for evaluating the predictive power of ML algorithms. The AUC value quantifies the overall ability of the ML algorithm to discriminate between the risk types of the patient. A random ML algorithm has an AUC value of 0.50 and a perfectly accurate ML algorithm has an AUC value of 1.0. The dotted line (Figure 5.7) has an AUC value of 0.50 and is the performance of a random classifier. A perfect ML algorithm would yield a point in the upper left corner or the coordinate (0, 1) of the ROC curve. Therefore, the ML algorithms are evaluated using the AUC value (higher the better), where AUCs greater than 0.80 are considered as good classifiers. Furthermore, the ROC curve, enables the hospital administration to make a business decision of choosing an appropriate threshold value to convert the probability value to either high-risk or low-risk patient. For instance, a threshold of 0.50 indicates that any value below 0.50 is low-risk and above 0.50 is high-risk. In most cases, the increase in accuracy of classifying one class is achieved at the expense of another class. Therefore, a ML algorithm may have any threshold value (e.g., 0.70), which depends on the objective

of the clinic (i.e., to predict low-risk patients accurately or high-risk patients accurately). Once the threshold value is set, the probability values below the threshold are categorized as low-risk patients and values above the threshold are categorized as high-risk patients.

5.2.4 Scheduling Rules

5.2.4.1 Current Practice

Hospitals schedule patients on a first call first appointment basis. Also, they use different overbooking strategies to compensate for no-shows. However, no specific approach has been universally adopted by the clinics. Most clinics use one of the two approaches described below:

- **Round-Robin**: Patients are added one by one to a slot in a sequential fashion starting from the first slot. Overbooking occurs only when the patient cannot be scheduled to an empty slot. Then the patients are overbooked in a sequential fashion starting from the first slot.

- **Evenly Distributed**: Similar to Round-Robin rule, the patients are scheduled one by one to a slot in a sequential manner starting from the first slot and overbooking occurs only when a patient cannot be single booked. However, unlike the Round-Robin rule, the overbooked slots are not sequential and are spaced at specific intervals across the clinic operating hours to avoid clinic overcrowding. In most cases, the hospital uses the average no-show rate to pre-specify the overbooked slots at specific intervals so that they are evenly distributed. For example, if two of the five slots are to be overbooked, then Slots 1 and 5 are overbooked so that they are evenly distributed across the schedule. However, this may not be appropriate in situations where a patient has to be scheduled for more than one slot. If a patient requires more than one slot, then the overbooked slots must be continuous. Therefore, in this paper, the distribution of the slots to be overbooked is modified and determined using the following procedure:

Step 1: Divide the total number of slots (S) for a given day into approximately three equal parts - Part 1, Part 2 and Part 3. The number of slots in Part 1 and Part 3 is the nearest integer of the total number of slots divided by 3 (i.e., $\left\lceil \frac{S}{3} \right\rceil$). The remaining slots is the number of slots in Part 2. For example, if there are 5 slots, then Part 1 and Part 3 has two slots each and Part 2 has only one slot. Thus, Slots 1 and 2 belong to the Part 1, Slot 3 belongs to the Part 2, and Slots 4 and 5 belongs to the Part 3.

Step 2: Determine whether the number of slots overbooked is less than the flat overbooking limit set by the clinic based on the average no-show rate. If YES, then proceed to step 3. Else, STOP as maximum overbooking limit has reached.

Step 3: Select the part that has the least number of overbooked slots. If two or more parts have the same number of overbooked slots, then randomly select one.

Step 4: If Part 1 or Part 2 is selected, then overbook the patient to the first available slot starting from the beginning of that part. If Part 3 is selected then overbook the patient to the first available slot starting from the end of that part.

Step 5: If the patient cannot be overbooked in that part, then try overbooking to the first available slots in the other parts.

Step 6: If the patient cannot be accommodated in any of the three parts, then the patient has to be scheduled for another day.

In addition, both the policies have a flat-overbooking percentage that depends on the average no-show rate of the clinic. The Round-Robin and Evenly Distributed rules are considered as baseline cases for comparison against other proposed rules in this chapter.

To illustrate the current practice,namely Round-Robin and Evenly Distributed scheduling rules, consider a situation in which a clinic has 5 slots of 15 minutes duration each, flat overbooking percentage of 60% (i.e., 3 out of 5 slots are overbooked), and a total of six sequential patients calls for appointment. The first two patients have an estimated service time of 30 minutes (i.e., requires 2 slots for service) and the remaining four patients have an estimated service time of 15 minutes. For the given situation, the schedules generated using

Round-Robin rule and Evenly Distributed rule are illustrated in Figure 5.8. For the Evenly Distributed rule, Part 1 has Slots 1 and 2, Part 2 has Slot 3, and Part 3 has Slots 4 and 5.

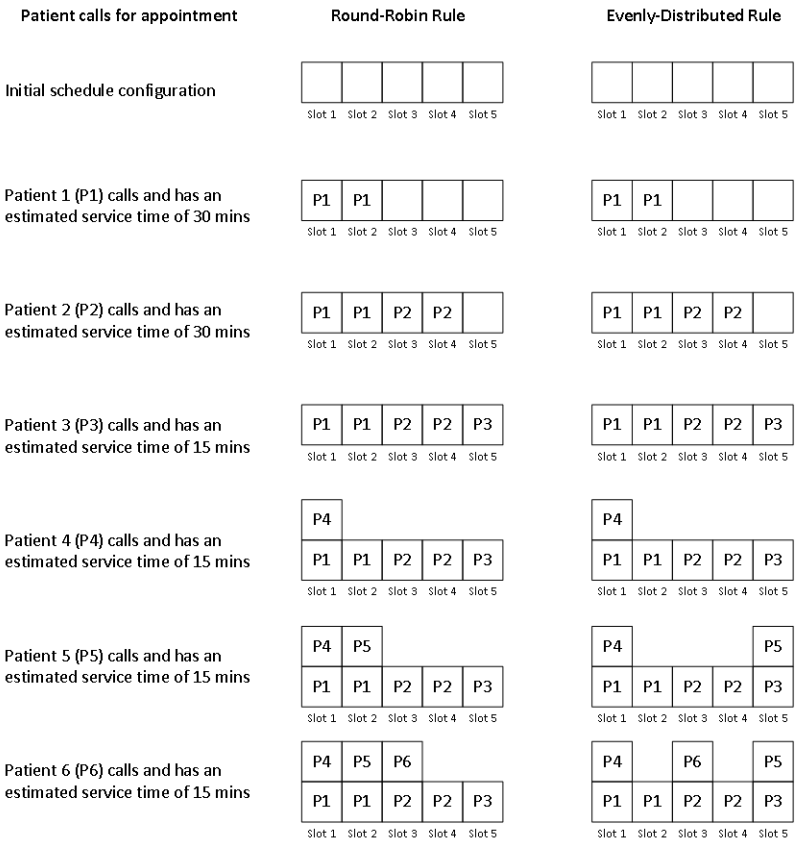


Figure 5.8: Illustration of Round Robin and Evenly Distributed Rules

5.2.4.2 Proposed Scheduling Rules

The proposed scheduling rules leverage the patient risk output obtained from the machine learning algorithm. The rules are a combination of two factors, namely, sequencing policy and overbooking policy. The sequencing policy determines the order in which a patient calling for an appointment is scheduled. The overbooking policy determines the type of patients who can be scheduled together in the same slot. In this chapter, four different sequencing policies and two overbooking policies are developed, resulting in a total of eight scheduling rules.

The sequencing and overbooking policies are developed based on the risk type of the

patients and the appointment duration assigned to each patient. The ML algorithm classifies the patient as high-risk(H), if the patient is likely to miss an appointment, and as low-risk (L), if the patient is likely to come for the appointment. The appointment duration of each patient varies depending on the patient type (new or return patient) and type of service required. The appointment duration can be classified as brief (B), intermediate (I) and extended (E). Based on our discussions with the different members of the clinical care team, it is evident that the clinic staff can estimate the appointment duration of a patient based on their knowledge and experience.

5.2.4.3 Sequencing Policies

As discussed in Section 5.2.4.1, the current practice of scheduling (Round-Robin and Evenly Distributed rules) does not use any sequencing policy and schedules the patients on first call first appointment basis in a sequential manner. To leverage the output obtained from the no-show prediction model discussed in Section 5.2.2, two sequencing policies based on the risk type of the patient and two more sequencing policies considering the variable appointment duration are proposed. Thus, four sequencing policies are proposed in this chapter to determine the scheduling sequence for a given day and are described below:

- **LRBG**: Schedule low-risk (L) patients to the first available slot starting from the beginning and high-risk (H) patients to the first available slot starting from the end of the clinic session (LLLL....HH).
- **HRBG**: Schedule high-risk (H) patients to the first available slot starting from the beginning and the low-risk (L) patients to the first available slot starting from the end of the clinic session (HH....LLLL).
- **EABG**: Schedule extended appointments (E) to the first available slot in the beginning, and brief (B) and intermediate (I) appointments to the first available slot starting from the end of the clinic session (EEE....BIIB).

- **BIBG**: Schedule brief (B) and intermediate (I) appointments to the first available slot in the beginning, and extended appointments (E) to the first available slot starting from the end of the clinic session (BIBB....EE).

5.2.4.4 Overbooking Policies

The sequencing rule tries to overbook only if the patient cannot be scheduled to an empty slot. A good overbooking policy must ensure that the probability of at least one patient coming on time for the appointment is high and the probability of both the patients coming on time for the appointments (i.e., risk of collisions) is low. Note that the patient may require more than one slot depending upon their appointment duration, and the multiple slots reserved for a patient must be continuous to ensure continuity of service. The two overbooking policies proposed in this chapter are the following:

- **OB1**: The OB1 policy overbooks by combining low-risk (L) and high-risk (H) patients to the same slot. Also, this policy overbooks the patient to the first available slot that meets the criterion of low-risk and high-risk combination.

- **OB2**: The OB2 policy also overbooks by combining low-risk (L) and high-risk (H) patients to the same slot. However, OB2 policy distributes the overbooked slots evenly across the schedule instead of overbooking the first available slot that meets the criterion of low-risk and high-risk combination. The procedure to distribute the overbooked slots is similar to the overbooking procedure of Evenly Distributed rule. The main difference is that the OB2 policy ensures the criterion of high-risk and low-risk combination while overbooking and does not enforce a flat overbooking percentage.

For example, consider a situation with 5 slots in which each slot has exactly one patient scheduled to it. Further, the patient in the first slot is high-risk, while the patient in the remaining slots are all low-risk (i.e., HLLLL). Since all slots are already single-booked, any additional patient calls for appointment must either be overbooked or scheduled for

another day. Suppose if two patients, one low-risk and one high-risk, call for appointment, then the schedule obtained based on overbooking policies, OB1 and OB2, is illustrated in Figure 5.9.

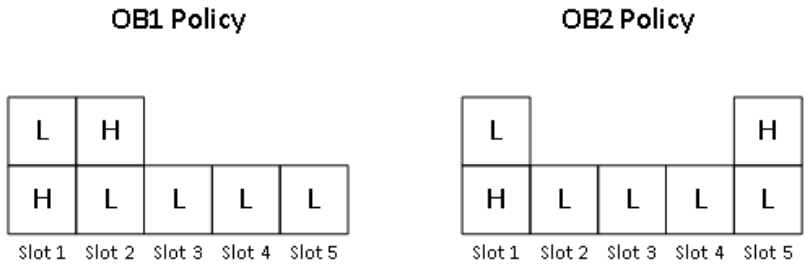


Figure 5.9: Illustration of OB1 and OB2 policies

If the next patient call is a low-risk patient, then the patient cannot be scheduled in any of the slots because it is not possible to meet the criterion of low-risk and high-risk patient combination. However, if the next patient call is a high-risk patient, then the patient will be scheduled in slot 3 for both OB1 and OB2 policy. Note that scheduling two low-risk patients to the same slot increases the probability of both the patients coming for the appointment and results in resource overburden and increased patient waiting time. Also, scheduling two high-risk patients to the same slot decreases the probability of at least one patient coming for the appointment and leads to increase in resource idle time. Hence, a combination of scheduling low-risk and high-risk patient to the same slot is alone studied.

5.2.4.5 Evaluating the Scheduling Rules

The primary objectives of the scheduling rules are to improve patient satisfaction and resource utilization. In this chapter, the average patient waiting time ($WAIT$) and the average number of patients unable to get an appointment for the day under consideration ($UNSCH$) are used as performance measures to quantify patient satisfaction. Similarly, the average resource idle time ($IDLE_r$) and average resource spillover time ($OVER_r$) are used to quantify utilization of resource r. Spillover time is the additional time spent by a resource in providing service after the scheduled appointment duration. Thus, the performance of

the scheduling rules are evaluated using the metrics that reflect patient satisfaction and resource utilization. The scheduling rules are evaluated in two ways:

1. **Total Cost Method**: Assign appropriate cost for each metric and obtain the total cost associated with the scheduling rule. Determine the best scheduling rule that minimizes the total cost.

2. **Multi-Criteria Decision Making Method**: Obtain relative weights associated with each performance metric from the hospital decision makers. Determine the overall score associated with each scheduling rule and rank them in order.

Equation (6.1) represents the total cost (TC) associated with the schedule based on the cost of patient's waiting time (C^A), unscheduled patient (C^U), resource's idle time (C_r^I), and resource's spillover time (C_r^O). Equation (5.9) represents the overall score (S) associated with the schedule based on the relative weight of patient's waiting time (W^A), unscheduled patient (W^U), resource's idle time (W_r^I), and resource's spillover time (W_r^O).

$$TC = [C^A \times WAIT] + [C^U \times UNSCH] + \sum_r [C_r^I \times IDLE_r] + \sum_r [C_r^O \times OVER_r] \tag{5.8}$$

$$S = [W^A \times WAIT] + [W^U \times UNSCH] + \sum_r [W_r^I \times IDLE_r] + \sum_r [W_r^O \times OVER_r] \tag{5.9}$$

In order to evaluate each of the 8 scheduling rules proposed in this chapter, the scheduling process is simulated, where patients call for an appointment in a sequential manner and are scheduled based on a particular scheduling rule. Figure 5.10 illustrates the scheduling process using a flow-chart. The schedule is generated by adding one patient at a time as and when they call for appointments without knowing the number of patients who may call at a later time and their associated risk of no-show. Further, the slot to which a patient is scheduled is determined by the scheduling rule under consideration. The service process is simulated in accordance with the fully constructed schedule and performance measures are computed to evaluate the schedule. Since the scheduling process is sequential, the fully

generated schedule mainly depends on the order in which the patient calls for appointments (i.e., patient call-in sequence). Thus, to ensure the robustness of the scheduling rules, different call-in sequences are generated using the same set of patients and the corresponding schedule is constructed based on the scheduling rule under consideration.

The schedules generated for all the call-in sequences (i.e., output of Figure 5.10) is given as the input to the service process (Figure 5.11), where the average performance measures associated with the schedules are determined. Figure 5.11 illustrates the simulation of the service process used to determine the performance measures. Given the schedule, the service process simulates the patient's arrival to clinic based on their risk (either show or no-show) and generates the service times of that patient for each phase.

The patient calls for appointments and their service times can be simulated from the fitted theoretical distribution using the observed data. However, patient no-show value cannot be simulated from a distribution. The ML algorithm provides the probability of a patient being a no-show (p) and also the risk type of the patient. Therefore, to simulate the patient no-show value, a uniform random variable (u) between 0 and 1 is generated, and compared with the patient no-show probability (p). If $u > p$, then the patient will arrive for the appointment, and if $u <= p$, then the patient will be a no-show. It is important to note that the schedule and the associated performance measures change depending on the sequence in which the patients call. Therefore, numerous call-in sequences are generated, and for each call-in sequence, the service process (patient show value and service time value) is replicated several times. Thus, the performance measures obtained is the average across all replications and call-in sequences.

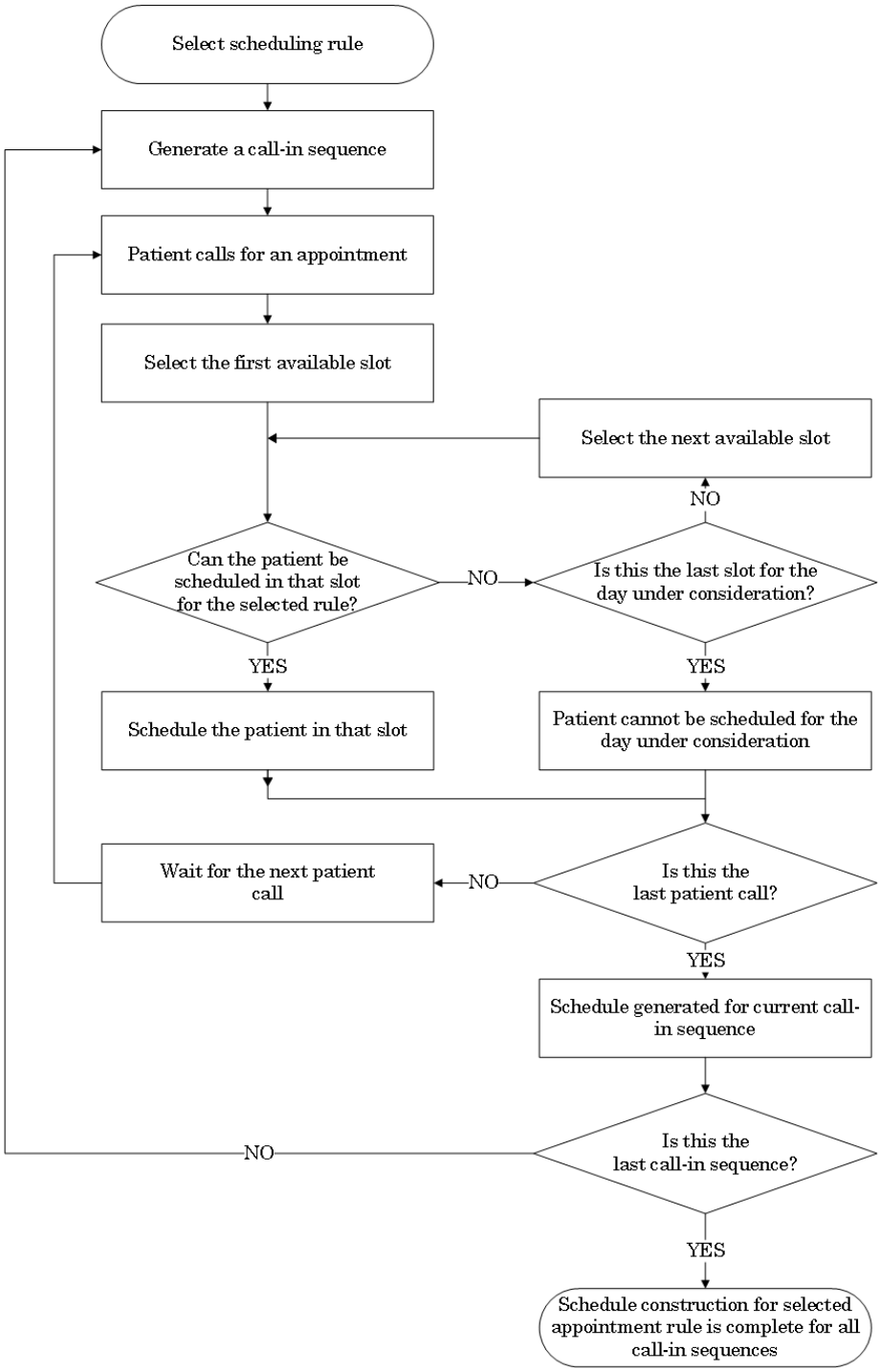


Figure 5.10: Flow chart illustrating patient scheduling process

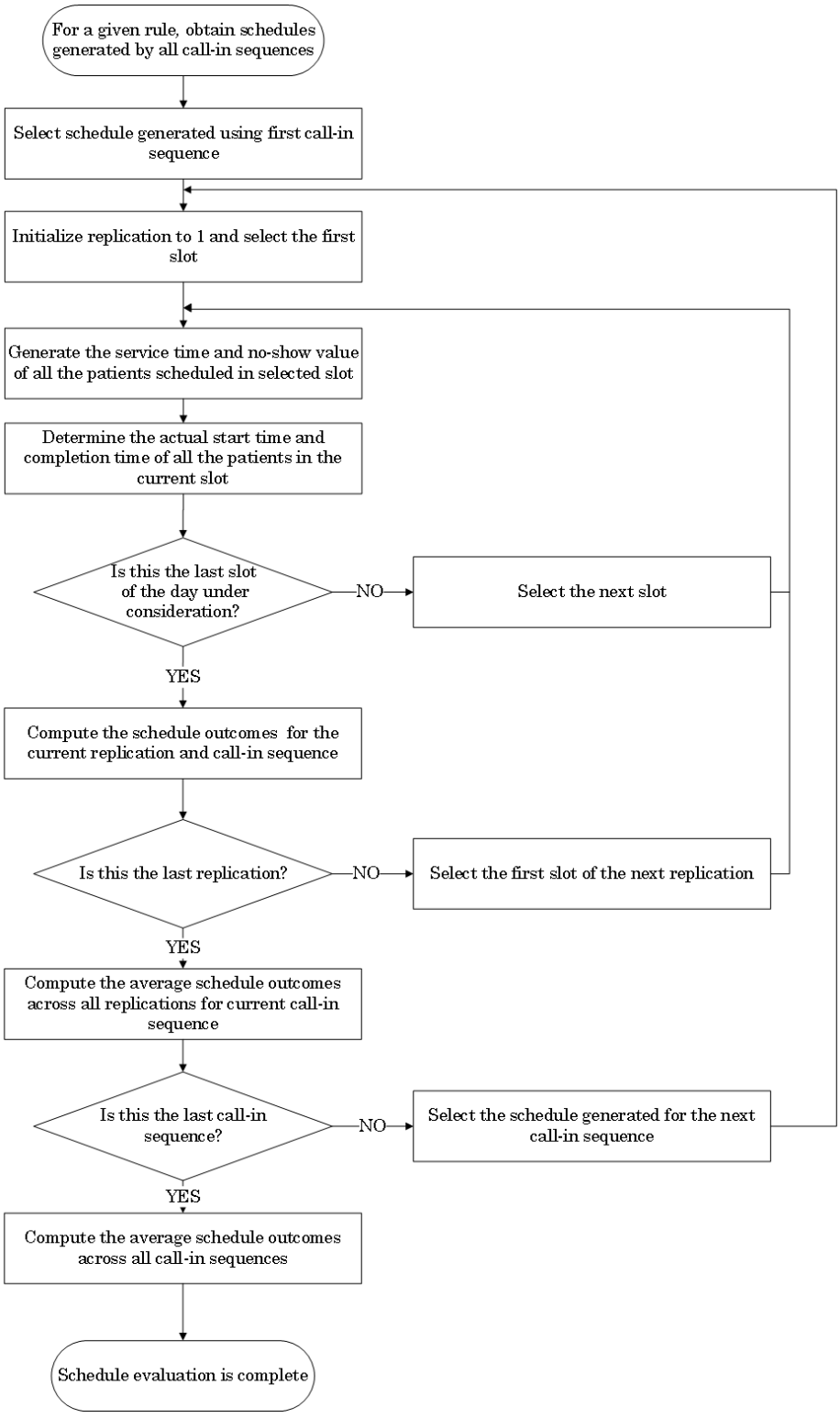


Figure 5.11: Flow chart illustrating schedule evaluation

5.3 Conclusions

Clinics overbook appointments to handle high patient demand and compensate for patient no-shows. However, the commonly adopted overbooking approaches, namely, Round-Robin and Evenly Distributed rules, assume all patients are equally likely to miss their appointments and do not consider the risk of both patients showing up for appointment when overbooked. Such practices lead to resource burnout, long patient waiting times in case of collisions and resource idle time in case of no-shows.

In this chapter, we propose a framework for scheduling patients. First, the patients are differentiated based on their likelihood of coming for their appointments (high-risk or low-risk). The data obtained from the various sources are used to classify the patient as either high risk or low risk of missing the appointment. This information is then used to make sequencing and overbooking decisions to generate the schedule. We proposed eight scheduling rules derived from four sequencing policies (LRBG - low risk patients in the beginning, HRBG - high risk patients in the beginning, EABG - extended appointments in the beginning, BIBG - brief and intermediate appointments in the beginning) and two overbooking policies (OB1 - overbook the first slot that meets the criterion of high risk and low risk patient combination, OB2 - distribute the overbooked across the schedule while ensuring the high risk and low risk patient combination). To evaluate the scheduling rules, six performance metrics, namely, nurse idle time, nurse spillover time, physician idle time, physician spillover time, patient waiting time and number of unscheduled patients, are used. Since the six performance metrics conflict with one another, two approaches are used to determine the best schedule: (1) use of relative costs associated with the performance metrics; (2) use of relative weights associated with the performance metrics. The application of the proposed scheduling framework is illustrated in Chapter 6 using a real case study.

Chapter 6

A Case Study for the Evaluation of the Proposed Policies for Scheduling Appointments

6.1 Introduction

The proposed scheduling framework in Chapter 5 is a three step procedure. First, the patient data is obtained from various sources such as electronic health records and Google maps API. In the second step, the data is cleansed by detecting and correcting inaccurate or missing records. The cleansed data is used as inputs to machine learning (ML) algorithms which classify (predict) the patient's no-show risk as low-risk (likely to come for the appointment) and high-risk (likely to miss the appointment). The best ML algorithm is then chosen for patient classification. Finally, different appointment scheduling rules that leverage the predictions obtained from the best ML algorithm are proposed and evaluated.

In this chapter, the proposed scheduling framework is demonstrated using a case study with real data obtained from a Family Medicine Clinic at Penn State Hershey Medical Center (PSHMC) located in Central Pennsylvania. PSHMC is a part of the Pennsylvania State University and the Penn State Health Care Partners clinically integrated network of regional care providers. It has over 1,100 physicians and treated over million outpatients in 2016. The data used in the chapter are approved under PSHMC institutional review board (IRB) *STUDY00004553: Design and Analysis of Appointment Systems for Outpatient*

Clinics.

6.2 Data Collection

The historical data was obtained from the scheduling system of the Family Medicine Clinic and included the electronic health records (EHR) of patient visits from September 2014 to August 2016. The data has a total of 76,285 patient visits, and for each patient visit included the following:

1. **Patient Information**: Age, Sex, Gender, Race, Marital Status, Zip Code, Insurance Group

2. **Appointment Information**: Time stamp of patient calls for appointment, timestamp of appointment date, timestamp of patient visits at each phase (i.e., service start time and completion time at each phase), appointment duration, appointment type

Some of the data obtained from the EHR can be directly used as inputs to ML algorithms for classifying patients based on their risk of no-show and for the simulation model for evaluating the proposed scheduling rules. However, some data require pre-processing to obtain meaningful inputs. Therefore, using the timestamp data, the following information is obtained for each patient visit:

1. **Appointment Delay**: The number of days between the call for appointment and appointment date

2. **Service Time at Each Phase**: The difference between the timestamp of service completion time and service start time of a phase will provide the service time for that phase (e.g., nurse phase, physician phase).

Similarly, using patient's and clinic's zip code, the following additional inputs are obtained for each patient visit, on the day of their appointment.

1. **Driving Time**: Extracted driving time (in minutes) data between the two zip codes using Google maps API

2. **Weather Information**: Extracted forecasted weather data, namely, minimum temperature, maximum temperature, precipitation probability, precipitation intensity, precipitation type, using the weather API of a commercial weather provider.

In addition, the following aggregate data for each day is obtained by descriptive analysis of the EHR data and is used to estimate the parameters for the simulation models that evaluate the scheduling rules:

1. Total number of patient calls per day

2. Number of Brief, Intermediate, Extended Appointments per day

6.3 Clinic Setting

The Family Medicine Clinic under study operates for 8 hours a day and 5 days a week. The entire operating hours of the clinic is divided into 15-minute slots, resulting in a total of 32 slots per day per physician. Depending on the type of procedure/treatment required, the clinic schedules the patient for 15 (brief), 30 (intermediate) or 45 (extended) minute appointments. Therefore, a patient having an expected service time of 15 minutes is assigned one slot and a patient having an expected service time of 45 minutes is assigned three continuous slots.

Further, the clinic has six physicians to provide service. However, on any given day, there are only two full time equivalent (FTE) physicians and two FTE nurses present, where each patient is first seen by a nurse and then by a physician. Moreover, the first one-third of the appointment duration is allocated for the nurse and the remaining two-thirds of the appointment duration is reserved for the physician. For example, if the appointment duration is 30 minutes, then the nurse is expected to serve the patient in the first 10 minutes and the physician is expected to serve the remaining 20 minutes. However, the actual

service time may vary, and the resource may complete their service earlier or later than the appointment end time. The clinic also experiences an average no-show rate of 30% and adopts overbooking to compensate for no-shows.

Based on the city in which the clinic is located, the median salary of full time equivalent (FTE) physicians is about $192,000/year, the median salary of FTE nurses is about $67,500/year, and the median salary of the patients (city population) is about $58,500/year. The hourly wage can be computed by dividing the annual salary by 2080 hours (8 hours/day $\times$ 5 days/week $\times$ 52 weeks/year = 2080 hours/year). Therefore, hourly wage of FTE physicians, nurses and patients are approximately $92, $32, and $28. Thus, the idle time cost of nurses (C_1^I) and physicians (C_2^I), and the waiting time cost (C^W) of patients is assumed to be equal to their hourly wage. Generally, the patients are expected to be served within the appointment duration. However, if a resource at a slot serves beyond the appointment end time of the slot, then it may lead to resource burnout, emotional stress, medical errors, decrease in productivity, hospital acquired infections, and early retirement. In this paper, the cost of resource spillover time is assumed to be 1.5 times the cost of resource idle time. Further, based on our interaction with the staff at the hospital, it is determined that the average net revenue of treating a patient is $40.

The aggregate data (total number of patient calls per day and number of Brief, Intermediate, Extended Appointments per day) and service time data obtained from EHR are used to fit a theoretical distribution to simulate the scheduling and the service process discussed in Chapter 5. Using a theoretical distribution, instead of an empirical distribution, provides greater flexibility to test cases beyond those observed in practice. A Kolmogorov-Smirnov test (at $\alpha = 0.05$) was used to identify the theoretical distributions that had a good fit to the observed data. The test results showed that Poisson distribution with a mean of 60 calls was a good fit for total patient calls for appointment. Also, Log-normal distribution was a good fit for nurse and physician service times for brief, intermediate and extended appointments. The mean and standard deviations of the log-normal distributions for service times along with a summary of other parameters used in the simulation model are shown in Table 6.1.

Table 6.1: Summary of model parameters for simulation model

Parameter	Value
Total patient calls for appointment	Poisson with rate = 35
Proportion of brief, intermediate and extended appointments	25%, 55%, 20%
Duration of brief, intermediate and extended appointments	15, 30 and 45 minutes
Average no-show rate	30%
Number of FTE nurses	2
Number of FTE physicians	2
Nurse Service time (in minutes) for brief appointments	Log-normal $\sim$ (5,0.75)
Nurse Service time (in minutes) for intermediate appointments	Log-normal $\sim$ (10,1.5)
Nurse Service time in (in minutes) for extended appointments	Log-normal $\sim$ (15,2.25)
Physician Service time (in minutes) for brief appointments	Log-normal $\sim$ (10,1.5)
Physician Service time (in minutes) for intermediate appointments	Log-normal $\sim$ (20,3)
Physician Service time in (in minutes) for extended appointments	Log-normal $\sim$ (30,4.5)
Total number of slots per day per physician	32 slots
Total number of call-in sequences	5000
Total number of replications	500
Costs/hour ($C^W, C_1^I, C_1^O, C_2^I, C_2^O$)	\$ 28, 32, 48, 92, 138
Cost of unscheduled patient (C^U)	\$40

6.4 General Observations from Data Analysis

Based on the analysis of data, some insights are obtained on the risk-type of the patient as a function of certain independent variables. Similar to prior research, it is found that longer appointment lead time and new patient types increase the risk of no-show. However, several new insights are also obtained. The analysis of the appointment information indicates that patients with appointments in the month of July or December, appointments in the morning session, or appointments with brief duration are likely to miss their appointments. As far as the weather information is considered, if maximum temperature is below 10 (°F) or if the

precipitation intensity is over 0.40 inches/hour, then the patient is likely to be a no-show.

6.5 Parameter Settings for Machine Learning (ML) Algorithms

As discussed in Table 5.1 of Chapter 5, the following information was retrieved from various sources for each patient visit and are used to train and test the machine learning algorithms:

1. **Patient Information**: Age, Sex, Marital Status, Race, Driving time, Insurance group, Patient type, Visit type
2. **Appointment Information**: Day, Month, Session, Duration, and Appointment delay
3. **Forecasted Weather Information**: Forecasted maximum temperature, Forecasted minimum temperature, Precipitation probability, Forecasted precipitation intensity, Precipitation Type

Thus, a total of 18 independent variables are used as inputs to the ML algorithms, while the patient show/no-show is the output. The data had a total of 76,285 patient visits, and is randomly partitioned into training and testing dataset. The training dataset is used to train the ML algorithm to uncover the underlying relationship between the inputs and the outputs, and the testing dataset is used to evaluate the trained ML algorithm. About two-thirds of the data (50,858 patient visits) is used for training the machine learning (ML) algorithms and the remaining one-third (25,427) is used for testing and evaluation.

Among the 5 ML algorithms evaluated, only Stacking requires the base-level and meta-level learning algorithms to be specified. Therefore, based on the review of literature, popular non-linear algorithms, namely, neural networks, random forests and gradient boosting machines are chosen as the base-level classifiers and a simple linear algorithm, namely, logistic regression, is chosen as the meta-level classifier. In this case study, the ML algorithm with the highest AUC value is considered as the best algorithm for classifying

the patients based on their risk of no-show. Table 6.2 provides a summary of the parameters settings for the ML algorithms.

Table 6.2: Summary of model parameters for machine learning algorithms

Parameter	**Value**
Number of patient visits in training data	50,858
Number of patient visits in testing data	25,427
Total number of independent variables	18
Number of classifiers evaluated	5
Base-level classifiers for Stacking	ANN, RF, SGBDT
Meta-level classifiers for Stacking	LR

6.6 Results of Machine Learning Algorithms

The caret package in R statistical computing software was used to train and test the ML algorithms (Kuhn, 2008). The experiments were conducted on a computer with 8GB RAM, Intel i5 2.50 GHz processor running Windows 10.

Training Phase

Out of the 50,858 patient visits in the training dataset, 13,770 patients are no-shows (high-risk), and 37,088 patients arrived for their appointments (low-risk). Further, to avoid the risk of overfitting in the training phase (i.e., model learning the noise in the data), a 5-fold cross validation is performed. In 5-fold cross validation, the training data set is split into 5-subsets, and one of the subset is set aside for validation and the remaining 4 subsets are used for training the ML algorithms. This process is repeated 5 times, where each of the 5 subsets are used exactly once for validation. For each ML algorithm, the cross validated AUC values are shown in Table 6.3.

Table 6.3: AUC values based on cross validation

ML Algorithm	Min.	Median	Mean	Max.
Logistic Regression	0.723	0.734	0.732	0.747
Artificial Neural Network	0.770	0.781	0.780	0.793
Random Forest	0.814	0.822	0.825	0.827
Gradient Boosting Machine	0.813	0.823	0.827	0.831
Stacking	0.851	0.856	0.858	0.868

Testing Phase

The trained ML algorithms are tested using the testing dataset. Out of the 25,427 patient visits in the testing dataset, 6,884 were high-risk patients, and 18,543 were low-risk patients. The best algorithm is chosen based on the AUC value. For each ML algorithm, AUC value obtained based on the testing dateset is shown in Table 6.4. It is evident that from cross-validation and testing that Stacking has the highest AUC value when compared to other methods, and therefore is chosen to classify the patients based on their risk type as high-risk or low-risk in this case study.

Table 6.4: AUC values based on testing datasets

ML Algorithm	AUC
Logistic Regression	0.717
Artificial Neural Network	0.774
Random Forest	0.812
Gradient Boosting Machine	0.816
Stacking	0.846

Table 6.5 shows the classification results from the Stacking algorithm on the testing dataset for a randomly chosen threshold value of 0.65. The Stacking algorithm correctly

classified 4,856 out of 6,884 high-risk patients, and 16,473 out of 18,453 low-risk patients. Using Equations (5.5) - (5.7) in Chapter 5 and the data in Table 6.5, the accuracy, sensitivity, and specificity of the Stacking algorithm were calculated as 83.88%, 70.54% and 88.83%, respectively.

Table 6.5: Predicted versus actual class with a threshold of 0.65

Predicted\Actual	High-Risk Patient	Low-Risk Patient
High-Risk Patient	4856	2070
Low-Risk Patient	2028	16473

Based on the analysis of the data, it is evident that Stacking algorithm is more suitable to classify the patients as high risk or low risk for the case study data. The trained Stacking classifier is then used to classify the patients in real-time. This information is used by all the 8 proposed scheduling rules to assign the patient to a slot.

6.7 Results of Scheduling Rules

In this section, the eight scheduling rules proposed in Chapter 5 are compared with two scheduling rules (Round Robin and Evenly Distributed) that are common in practice. The proposed scheduling rules are as follows:

1. LRBG + OB1
2. LRBG + OB2
3. HRBG + OB1
4. HRBG + OB2
5. BIBG + OB1
6. BIBG + OB2

7. EABG + OB1

8. EABG + OB2

(Refer to Section 5.2.4.3 and 5.2.4.4 for a description of the proposed rules)

The schedule is generated by adding one patient at a time as and when they call for appointments, without knowing the number of patients who may call at a later time and their associated risk of no-show. Further, the slot to which a patient is scheduled is determined by the scheduling rule under consideration. The service process is simulated in accordance with the fully constructed schedule and the performance measures are computed to evaluate the schedule. A simulation model was developed to evaluate each of the 10 scheduling rules (2 current + 8 proposed), using Microsoft Visual C++. Six different performance metrics, namely, nurse idle time, nurse spillover time, physician idle time, physician spillover time, patient waiting time, and number of unscheduled patients are used to evaluate the scheduling rules. Since the metrics are conflicting in nature, two different approaches are used to determine the final schedule.

1. **Total Cost Method**: Assign appropriate cost for each metric and obtain the total cost associated with the scheduling rule. Choose the schedule with the lowest cost.

2. **Multi-Criteria Decision Making (MCDM) Method**: Obtain relative weights associated with each performance metric from the hospital decision makers. Obtain the overall score associated with each scheduling rule and rank them in order.

6.7.1 Total Cost Method

Figures 6.1 shows the average total cost and its 95% confidence interval and Figure 6.2 is a box plot of the average total cost over 5000 sequences and 500 replications. Table 6.6 presents the average (Avg.) and standard deviation (STD) of total cost obtained over 5000 patient call-in sequences and 500 replications for different scheduling rules under study.

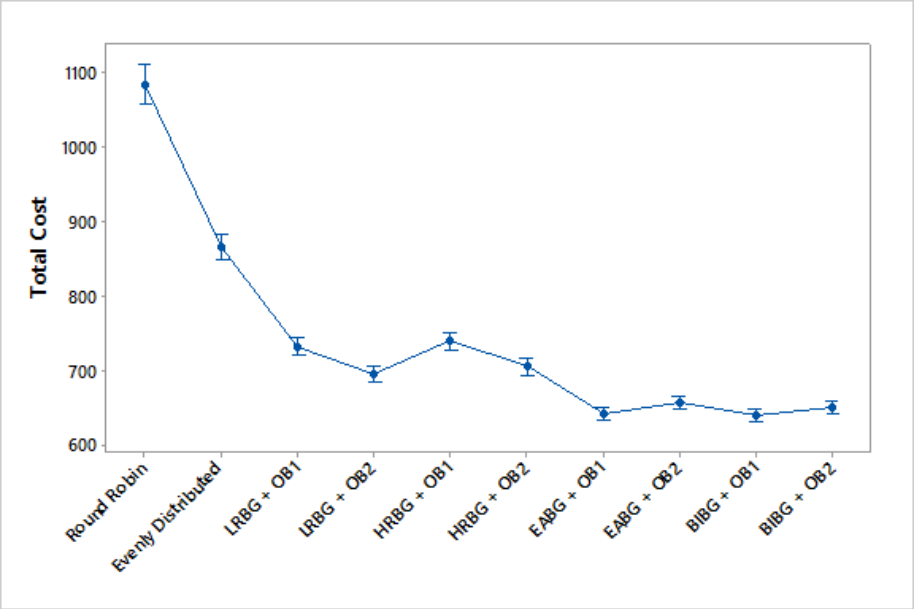


Figure 6.1: Average total cost for different scheduling rules

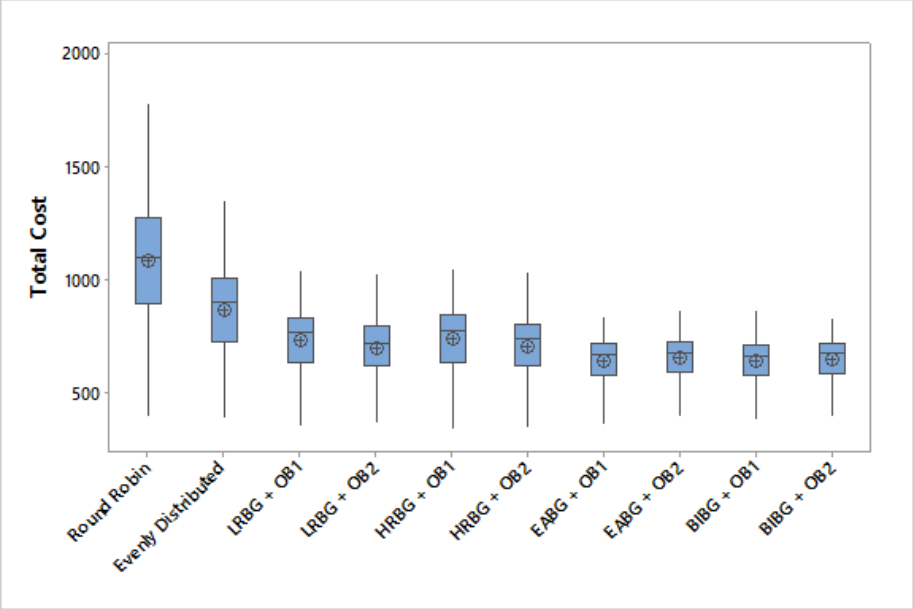


Figure 6.2: Plot of mean total cost for different scheduling rules

Table 6.6: Total cost for various scheduling rules over 5000 call-in sequences and 500 replications

Scheduling Rule	**Total Cost**	
	Avg.	**STD**
Round Robin	1084.3	298.4
Evenly Distributed	866.4	190.6
LRBG + OB1	732.2	139.4
LRBG + OB2	695.8	123.9
HRBG + OB1	739.2	140.7
HRBG + OB2	704.9	130.2
EABG + OB1	641.8	105.9
EABG + OB2	656.2	98.0
BIBG + OB1	639.8	101.1
BIBG + OB2	650.1	93.2

All the proposed scheduling rules that use the risk-type of patients to generate the schedule results in lower cost and lower standard deviation compared to the currently used rules (round-robin and evenly distributed rule) that do not consider any risk type to generate the schedule.

A Tukey multiple comparison test with a confidence coefficient of 0.95 is performed to test the statistical significance of the average total cost of the scheduling rules. Since there are 10 scheduling rules, a total of 45 pairwise comparisons (i.e., ${}^{10}C_2$ combinations) are possible. The test results indicate that all the proposed scheduling rules have a significantly lower average total cost compared to the current practice (Round Robin policy or Evenly Distributed rule). As indicated in Table 6.6, the average total costs are not significantly different from each other for the following 9 pairs of scheduling rules: (HRBG + OB1, LRBG + OB1); (HRBG + OB2, LRBG + OB1); (HRBG + OB2, LRBG + OB2); (EABG + OB1, EABG + OB2); (BIBG + OB1, EABG + OB1); (BIBG + OB2, EABG + OB1);

(BIBG + OB1, EABG + OB2); (BIBG + OB2, EABG + OB2); (BIBG + OB1, BIBG + OB2). The average total cost of all the remaining pairs of scheduling rules are significantly different.

The BIBG + OB1 rule results in the least total cost. However, based on statistical testing, the average total cost of BIBG + OB1, BIBG + OB2, EABG + OB1, and EABG + OB2 are not significantly different from each other, and therefore, these four rules can be considered as the best scheduling rules for the given hospital cost setting. *In other words, sequencing based on expected service times and overbooking based on the risk of patient no-shows appears to be better than sequencing and overbooking based on the risk of patient no-shows only.*

As discussed in Section 5.2.4.5, the total cost is obtained by using an appropriate cost (given in Section 6.3) for each of the six metrics. Table 6.7 shows the average and standard deviation values of each of the six performance measures, for all the scheduling rules under study. In addition to the statistical comparison of the total cost, a Tukey test is also performed to test the statistical significance of the six performance measures. It is observed that the physician idle time, physician spillover time, nurse idle time, nurse spillover time and patient waiting time of the proposed scheduling rules are significantly lower compared to the baseline cases - Round Robin and Evenly Distributed rules. Even though the baseline cases have less unscheduled patients compared to the proposed rules, it is not statistically significant when compared to LRBG + OB1, LRBG + OB2, HRBG + OB1, or HRBG + OB2. It is also important to note that no single scheduling rule achieves the best value with respect to all the six performance measures. For example, BIBG + OB1 results in the least nurse idle time, while EABG + OB1 results in the least physician idle time.

Table 6.7: Average and standard deviation of performance measures for various scheduling rules over 5000 call-in sequences and 500 replications

Scheduling Rule	NIDLE		PIDLE		NOVER		POVER		WAIT		UNSCH	
	Avg.	STD	Avg.	STD	Avg.	STD	Avg.	STD	Avg.	STD	Avg.	STD
Round Robin	61.5	10.6	74.4	6.0	17.5	3.6	98.2	46.1	9.1	3.9	2.7	0.6
Evenly Distributed	60.4	10.1	81.3	6.9	17.7	3.7	59.0	27.5	6.5	2.8	2.7	0.6
LRBG + OB1	55.7	8.8	67.6	6.2	11.5	1.7	42.9	19.3	5.5	2.3	3.2	0.7
LRBG + OB2	55.4	8.8	70.6	6.3	12.3	1.6	35.0	16.2	5.2	2.1	3.2	0.7
HRBG + OB1	56.9	9.3	68.3	6.5	11.9	1.8	42.6	19.2	5.7	2.4	3.2	0.7
HRBG + OB2	55.9	8.9	68.5	6.2	12.4	1.6	37.3	17.5	5.4	2.3	3.2	0.7
EABG + OB1	42.4	8.3	61.2	5.9	14.3	1.9	29.5	13.2	4.9	2.0	3.6	0.7
EABG + OB2	45.6	8.5	69.4	6.9	13.2	1.9	24.9	12.4	4.5	1.9	4.0	0.8
BIBG + OB1	42.3	8.5	64.3	6.7	15.0	1.9	27.3	12.5	4.8	1.9	3.6	0.7
BIBG + OB2	47.2	8.3	67.5	7.1	12.8	1.9	25.1	11.5	4.5	1.8	4.0	0.8

NIDLE - Nurse idle time; PIDLE - Physician idle time; NOVER - Nurse spillover time;

POVER - Physician spillover time; WAIT - Patient waiting time; UNSCH - No. of unscheduled patients

All times are in minutes

Based on the values in Table 6.7, it is clear that the nurse idle time, physician overload time and patient waiting time are always lower for sequencing based on expected service times (EABG and BIBG) compared to sequencing based on the risk of patient no-show (LRBG and HRBG), irrespective of the overbooking policies used. Similarly, irrespective of the overbooking policy, the number of unscheduled patients and nurse overload time are always lower for sequencing based on risk of patient no-show. Also, it is observed that the OB1 policy results in lower physician idle time, while the OB2 policy results in lower physician spillover time and patient waiting time. However, the overbooking policies do not have a substantial impact on nurse idle time or nurse spillover time.

Figure 6.3 is a value path graph that shows a graphical representation of the trade-off among the performance measures for different scheduling rules. To construct the graph, the ideal value for each performance measure is determined (i.e., the lowest value achieved for that performance measure average) and the values in Table 6.7 is scaled between 0 and 1 by

dividing the ideal value with the performance measure. Since all the performance measures are to be minimized, the ideal value is one and all others are less than one. Therefore, after scaling, a higher value of a scaled performance measure is better. The scaled values are plotted in the graph on y-axis with the performance measure in the x-axis.

The value path graph helps to visualize whether a scheduling rule dominates another. If a line has all its value higher than another line, then the line above dominates the line below. On the other hand, if the lines intersect, then they do not dominate each other. It can be seen from Figure 6.3 that all the lines intersect with each other, thereby indicating that none of the scheduling rules dominates another. In addition, the value path graph can also be used by the hospital administrators to make trade-offs and decide the scheduling rule that works best for their hospital. For example, consider a situation where the hospital adminstrators are interested in minimizing the overall idle time of the resources (i.e., both physician and nurse). The value path graph indicates that BIBG+OB1 rule is the best to minimize nurse idle time while EABG + OB1 is the best to minimize physician idle time. In addition, it also shows that if we choose EABG+OB1, then by sacrificing a small value in nurse idle time (when compared to BIBG+OB1), a substantial increase in physician idle time can be achieved. Therefore, to minimize the overall idle time, EABG+OB1 appears to be the best rule.

Figure 6.3: Value path graph illustrating trade-off among performance measures

6.7.2 Sensitivity Analysis

The evaluation of all the scheduling rules is based on a clinic experiencing an average no-show rate of 30% with a coefficient of variation (CV) of service time equal to 0.15. These values are based on the analysis of historical data set obtained from the electronic health records for a particular outpatient clinic. In order to gain some insights on the sensitivity of the proposed scheduling rules and test its applicability for other clinic settings, different values of average no-show rates and CV of service times are tested. We chose to vary the average no-show rates and CV of service time for two reasons: (i) the LRBG and HRBG sequencing rule, and the OB1 and OB2 overbooking policy are dependent on the no-show rate because it schedules the patients based on their likelihood of being a no-show. (ii) the EABG and BIBG sequencing policy schedules patients based on their expected service time. Thus, we analyzed the performance of the scheduling rules by varying the most sensitive parameters, namely, no-show rates and CV of service time.

6.7.2.1 Impact of Patient No-show Rate

Outpatient departments may experience high no-show rates of up to 34% (Geraghty et al. 2007, Dreiher et al. 2008). Therefore, average no-show rate is varied between 10% and 40% in increments of 10% to analyze its impact on the baseline cases and proposed scheduling rules. For each scheduling rule under study, Figure 6.4 shows the average total cost and its 95% confidence interval over 5000 sequences and 500 replications for different levels of patient no-show rates. It is evident that the proposed scheduling rules are always superior to the current rules (Round Robin and Evenly Distributed), irrespective of the average no-show rates of the clinic.

(a) (b)

(c) (d)

Figure 6.4: Average total cost for different levels of patient no-show rates

Table 6.8 presents the average (Avg.) and standard deviation (STD) of the total cost associated with each scheduling rule for different levels of no-show rates. Both the average total cost and its standard deviation are lower for all the proposed scheduling rules compared to the current practice. A Tukey test indicated that the average total cost of the proposed scheduling rules are significantly lower than the current rules for all levels of the no-show rates under consideration. Thus, the findings are consistent with the analysis in Section 6.7.1, that indicate that the proposed rules outperform the current rules used in the clinics, for all levels of no-show rates under study.

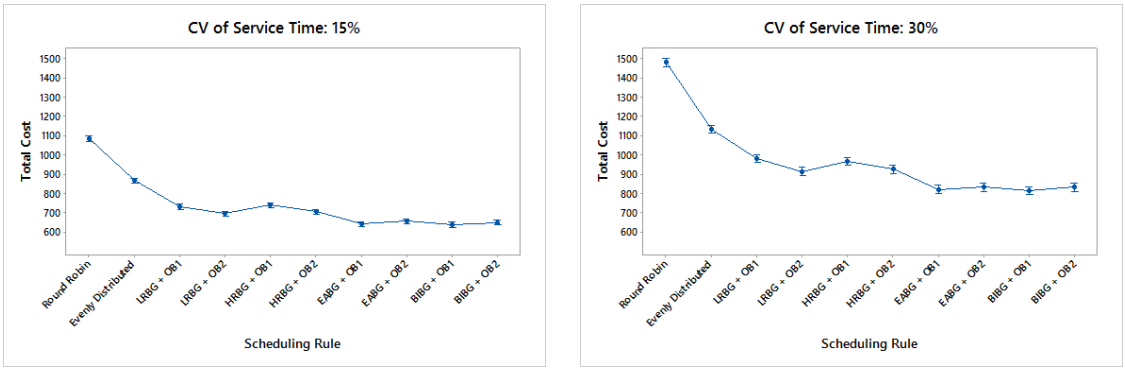
Table 6.8: Average total cost and its standard deviation for different levels of no-show rates

Scheduling Rule	Total Cost							
	No-show: 10%		No-show: 20%		No-show: 30%		No-show: 40%	
	Avg.	STD	Avg.	STD	Avg.	STD	Avg.	STD
Round Robin	947.8	92.6	1032.6	202.9	1084.3	298.4	967.9	286.0
Evenly Distributed	905.8	76.2	993.9	120.6	866.4	190.6	830.5	211.8
LRBG + OB1	804.7	57.5	751.6	87.3	732.2	139.4	676.3	162.0
LRBG + OB2	807.3	56.3	718.3	79.4	695.8	123.9	633.9	136.6
HRBG + OB1	809.2	57.2	768.7	86.0	739.2	140.7	684.2	164.9
HRBG + OB2	811.0	57.5	719.4	80.4	704.9	130.2	643.9	145.0
EABG + OB1	772.5	49.2	670.3	71.1	641.8	105.9	573.7	139.6
EABG + OB2	775.6	47.9	673.3	66.0	656.2	98.0	609.0	119.9
BIBG + OB1	773.0	48.3	671.6	69.1	639.8	101.1	576.5	134.0
BIBG + OB2	777.2	48.2	675.4	65.9	650.1	93.2	605.3	119.0

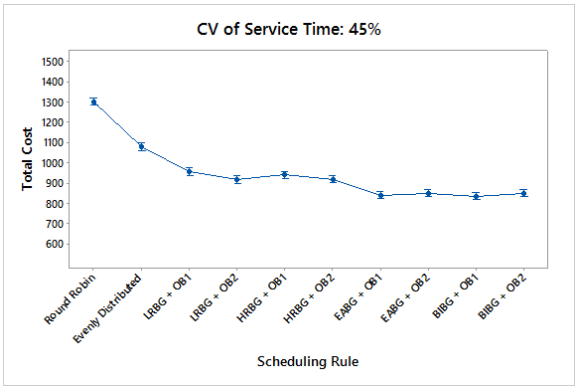
Also, it can be observed from Table 6.8 that the average total cost decreases as the no-show rate increases for all the eight proposed scheduling rules. Therefore, it can be concluded that the proposed scheduling rules perform better as the no-show rate increases. However, due to random overbooking approach, the average total cost of the current practice fluctuates as no-show rate increases.

6.7.2.2 Impact of Coefficient of Variation (CV) of Service Time

The CV of service time was set to 0.15 in prior analysis. To test its impact on the scheduling rules, two additional levels of service time CV of 0.30 and 0.45, are evaluated. Figure 6.5 shows the average total cost and its 95% confidence interval over 5000 sequences and 500 replications for different levels of service time CV. Even for different levels of service time CV, the proposed rules perform better with respect to average total cost and its standard deviation, and this is also confirmed by statistical significance testing.



(a) (b)



(c)

Figure 6.5: Average total cost for different levels of service time CV

Table 6.9 presents the average (Avg.) and standard deviation (STD) of the total cost associated with each scheduling rule for different levels of service time CV. It can be observed that BIBG + OB1 results in the least total cost for all levels of service time CV.

Table 6.9: Total cost of scheduling rules for different levels of CV of service times

Scheduling Rule	Total Cost					
	CV = 0.15		CV = 0.30		CV = 0.45	
	Avg.	STD	Avg.	STD	Avg.	STD
Round Robin	1084.3	298.4	1479.5	467.8	1301.0	385.2
Evenly Distributed	866.4	190.6	1132.5	286.3	1078.1	252.7
LRBG + OB1	732.2	139.4	982.42	228.1	957.2	202.2
LRBG + OB2	695.8	123.9	914.94	195.1	917.0	176.6
HRBG + OB1	739.2	140.7	967.53	224.2	940.1	197.5
HRBG + OB2	704.9	130.2	925.84	203.1	919.9	185.0
EABG + OB1	641.8	105.9	821.13	151.4	841.7	144.2
EABG + OB2	656.2	98.0	832.48	149.5	850.6	134.7
BIBG + OB1	639.8	101.1	815.86	145.8	836.3	137.7
BIBG + OB2	650.1	93.2	833.15	143.8	851.5	132.9

It can also be observed from Table 6.9 that the average total costs increases as the service time CV increases, particularly for scheduling rules that sequences based on expected service time of the patient (i.e., EABG + OB1, EABG + OB2, BIBG + OB1, BIBG + OB2). For all other scheduling rules, the average total cost fluctuates as the service time CV increases.

Also, based on Tables 6.8 and 6.9, it is observed that OB2 policy results in lower total cost for sequencing based on risk type of the patient, while OB1 policy results in significantly lower cost for sequencing based on expected patient service time.

6.7.3 Multi-Criteria Decision Making (MCDM) Method

Until now, the costs associated with the performance measures are used to combine them to a single total cost. In Section 6.3, we described how the cost of various performance measures were computed for this case study. In general, estimating the cost of some of the

measures, particularly cost of idle time of nurses, doctors and patient waiting times, are difficult. An alternative approach is to assign relative weights to the performance measures, using multi-criteria decision making. The relative weights represent the importance of the performance measure to the decision makers at a clinic. As shown in Figure 6.6, there are two main criteria - resource utilization and patient satisfaction. Under resource utilization, there are four sub-criteria - physician overload time, physician idle time, nurse overload time and nurse idle time. Under patient satisfaction there are two sub-criteria - patient waiting time and number of unscheduled patients.

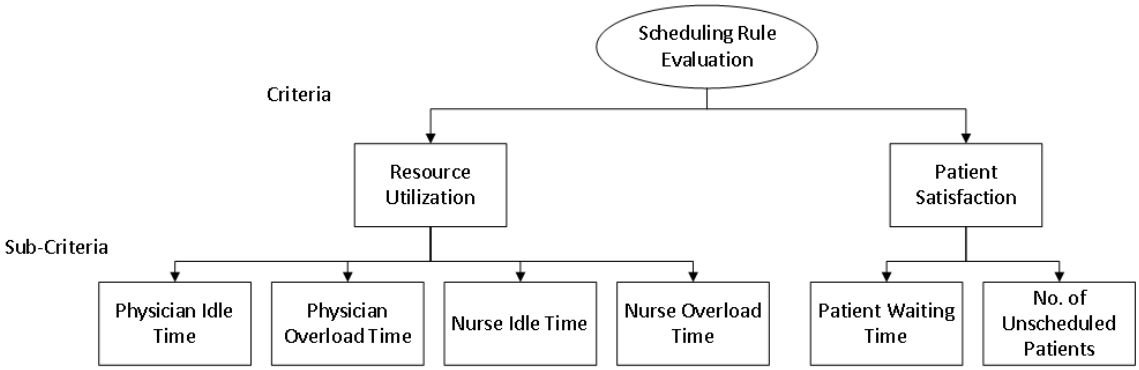


Figure 6.6: Schedule evaluation criteria and sub-criteria

The relative weights of the criteria and sub-criteria can be obtained from the decision makers (DM), using multi-criteria decision making (MCDM) methods. In some cases, only a single decision maker, such as the hospital administrator, may be involved in the decision making process; in other cases, multiple decision makers, including physicians, nurses, and hospital administrators, may be involved in the decision making process. The different MCDM methods to obtain the relative weights of the performance measures are discussed in Ravindran (2016). We will illustrate the following methods next:

1. Rating Method

2. Borda Method

6.7.3.1 Rating Method

In the Rating method, the DM gives a rating (r_j) for each criterion or sub-criterion on a scale of 1 (least important) - 10 (most important). Next, the weights of the criteria or sub-criteria are obtained by normalizing the ratings. Note that the DM can give the same rating to more than one criterion. Assuming K criteria, the weights are calculated as follows:

$$W_k = \frac{r_k}{\sum_{k=1}^{K} r_k} \quad \forall\, k = 1, 2, ..., K \tag{6.1}$$

Since the weights are obtained by normalizing, the summation of the the criteria weights will be equal to 1 (i.e., $\sum_{k=1}^{K} W_k = 1$). In the case of multiple DMs, the final rating for a criterion is the average of the ratings given by all the DMs for that criterion. The weights are then obtained by normalizing the final rating. To illustrate the rating method with single DM, let us consider different scenarios for the hospital. The scenarios represent the relative importance of the performance measures to a hospital.

Scenario 1

Consider a hospital in which the utilization of resources is of utmost importance compared to patient satisfaction. Further, the physician's time is considered more valuable than the nurse's time. It is also assumed that patients give more importance to clinic accessibility (i.e., the ability to get an appointment) than waiting at the clinic. An example of the DM ratings of the criteria for such hospitals and the corresponding weights obtained by the Rating method are shown in Table 6.10.

Table 6.10: Criteria ratings and their weights using the Rating method for Scenario 1

Criteria	DM Rating	Weights
Resource Utilization	8	0.80
Patient Satisfaction	2	0.20

The sub-criteria weights are also computed using the Rating method by adopting the

same procedure used to compute the main criteria weights. The DM provides a rating for the sub-criteria with respect to their main criterion. Tables 6.11 and 6.12 shows examples of the DM ratings for the sub-criteria and their weights, with respect to resource utilization and patient satisfaction, respectively.

Table 6.11: Sub-criteria ratings and their weights for resource utilization metrics by using the Rating method for Scenario 1

Resource Utilization Metrics	**DM Rating**	**Weights**
Physician Overload Time	9	0.36
Physician Idle Time	8	0.32
Nurse Overload Time	5	0.20
Nurse Idle Time	3	0.12

Table 6.12: Sub-criteria ratings and their weights for patient satisfaction metrics by using the Rating method for Scenario 1

Patient Satisfaction Metrics	**DM Rating**	**Weights**
Patient Waiting Time	3	0.25
No. of Unscheduled Patients	9	0.75

The last step is to obtain the final weights of the performance measures, which is the product of sub-criteria weights and its corresponding criteria weight. Table 6.13 shows the final weights of the performance measures obtained using the Rating method for Scenario 1. Note that the final weights are the products of the main criteria weights and their sub-criteria weights.

Table 6.13: Final weights of the performance measures obtained using the Rating Method for Scenario 1

Criteria (Criteria Weight)	**Sub-Criteria**	**Sub-Criteria Weight**	**Final Weight**
Resource Utilization (0.8)	Physician Overload Time	0.36	0.288
	Physician Idle Time	0.32	0.256
	Nurse Overload Time	0.20	0.160
	Nurse Idle Time	0.12	0.096
Patient Satisfaction (0.2)	Patient Waiting Time	0.25	0.050
	No. of Unscheduled Patients	0.75	0.150

Scenario 2

Some hospitals may emphasize patient satisfaction more than resource utilization. Among sub-criteria associated with resource utilization, the hospital may give more importance to resource overloading because it leads to resource burnout, stress and poor quality of patient care. Among patient satisfaction sub-criteria, it is assumed that patient's time is more important than accessibility. Tables 6.14 - 6.16 illustrate the DM's ratings for such hospitals and the corresponding weights obtained using the Rating method. The final weights of the performance measures for Scenario 2 are given in Table 6.17.

Table 6.14: Criteria ratings and their weights using the Rating method for Scenario 2

Criteria	**DM Rating**	**Weights**
Resource Utilization	2	0.20
Patient Satisfaction	8	0.80

Table 6.15: Sub-criteria ratings and their weights for resource utilization metrics by using the Rating method for Scenario 2

Resource Utilization Metrics	**DM Rating**	**Weights**
Physician Overload Time	10	0.33
Physician Idle Time	6	0.20
Nurse Overload Time	9	0.30
Nurse Idle Time	5	0.17

Table 6.16: Sub-criteria ratings and their weights for patient satisfaction metrics by using the Rating method for Scenario 2

Patient Satisfaction Metrics	**DM Rating**	**Weights**
Patient Waiting Time	9	0.75
No. of Unscheduled Patients	3	0.25

Table 6.17: Final weights of the performance measures obtained using the Rating Method for Scenario 2

Criteria (Criteria Weight)	**Sub Criteria**	**Sub-Criteria Weight**	**Final Weight**
Resource Utilization (0.2)	Physician Overload Time	0.33	0.066
	Physician Idle Time	0.20	0.040
	Nurse Overload Time	0.30	0.060
	Nurse Idle Time	0.17	0.034
Patient Satisfaction (0.4)	Patient Waiting Time	0.75	0.600
	No. of Unscheduled Patients	0.25	0.200

Scenario 3

In Scenario 3, we consider a hospital that considers both resource utilization and patient satisfaction as equally important. Further, all the sub-criteria are also given equal importance. Therefore, the DM is neutral and will provide the same rating for the criteria and

sub-criteria. Table 6.18 - 6.20 illustrate the DM's ratings and weights obtained using the Rating method for a Neutral DM. The final weights for the performance measures obtained using the Rating method for Scenario 3 is shown in Table 6.21

Table 6.18: Criteria ratings and their weights using the Rating method for for Scenario 3

Criteria	**DM Rating**	**Weights**
Resource Utilization	8	0.50
Patient Satisfaction	8	0.50

Table 6.19: DM rating and sub-criteria weights for resource utilization metrics using the Rating method for Scenario 3

Resource Utilization Metrics	**DM Rating**	**Weights**
Physician Overload Time	8	0.25
Physician Idle Time	8	0.25
Nurse Overload Time	8	0.25
Nurse Idle Time	8	0.25

Table 6.20: DM rating and sub-criteria weights for patient satisfaction metrics using the Rating method for Scenario 3

Patient Satisfaction Metrics	**DM Rating**	**Weights**
Patient Waiting Time	7	0.50
No. of Unscheduled Patients	7	0.50

Table 6.21: Final weights of the performance measures obtained using the Rating Method for Scenario 3

Criteria (Criteria Weight)	**Sub Criteria**	**Sub-Criteria Weight**	**Final Weight**
Resource Utilization (0.5)	Physician Overload Time	0.25	0.125
	Physician Idle Time	0.25	0.125
	Nurse Overload Time	0.25	0.125
	Nurse Idle Time	0.25	0.125
Patient Satisfaction (0.5)	Patient Waiting Time	0.50	0.250
	No. of Unscheduled Patients	0.50	0.250

6.7.3.2 Borda Method

Unlike the Rating method, the Borda method requires the ranking of the criteria from the most important to the least important from the DM. Then, points are assigned to each criterion based on their rankings. If there are K criteria, then the most important criterion gets K points, the next most important criterion gets $K-1$ points, and the least important criterion gets 1 point. Finally, the criteria weights are obtained by normalizing the points.

When the number of criteria is large, it is difficult for the DM to rank the criteria from most important to least important. In such situations, the Borda method also uses pairwise comparison of the criteria to determine the overall ranking. The pairwise comparison procedure asks for the DMs relative preference by presenting two criteria at a time. Hence, if there are K criteria, then the DM provides his/her preference for $\frac{K(K-1)}{2}$ pairwise comparisons. Based on the DMs responses, the ranking for the criteria are determined, and are then normalized to determine the criteria weights.

In case of multiple DMs, each DM independently provides the ranking of the criteria. Then, each criterion is given an overall point by analyzing the rankings assigned by the DMs for that particular criterion. For example, consider a situation with 3 DMs and 4 criteria (A, B, C, and D). In the Borda method with 4 criteria, criterion ranked 1 gets 4

points, criterion ranked 2 gets 3 points and so on. If two DMs assign rank 1 to criterion A and one DM assigns rank 3 to criterion A, then the overall points for criterion A is 10 (i.e., 2 DMs $\times$ 4 points + 1 DM $\times$ 2 points = 10). Let us consider two scenarios to illustrate the Borda method with a single DM.

Scenario 4

In Scenario 4, we consider a hospital similar to Scenario 1. However, we use the Borda method to determine the weights for such a hospital setting. Tables 6.22 - 6.24 illustrate the DM's rankings and the corresponding weights for the criteria and sub-criteria. Table 6.25 shows the final weights of the performance measures obtained using the Borda method for Scenario 4.

Table 6.22: Criteria rankings and their weights using the Borda method for for Scenario 4

Criteria	DM Ranking	Points	Weights
Resource Utilization	1	2	0.67
Patient Satisfaction	2	1	0.33

Table 6.23: DM ranking and sub-criteria weights for resource utilization metrics using the Borda method for Scenario 4

Resource Utilization Metrics	DM Ranking	Ranking Points	Weights
Physician Overload Time	1	4	0.40
Physician Idle Time	2	3	0.30
Nurse Overload Time	3	2	0.20
Nurse Idle Time	4	1	0.10

Table 6.24: DM ranking and sub-criteria weights for patient satisfaction metrics using the Borda method for Scenario 4

Patient Satisfaction Metrics	**DM Ranking**	**Ranking Points**	**Weights**
Patient Waiting Time	2	1	0.33
No. of Unscheduled Patients	1	2	0.67

Table 6.25: Final weights of the performance measures obtained using the Borda Method for Scenario 4

Criteria (Criteria Weight)	**Sub Criteria**	**Sub-Criteria Weight**	**Final Weight**
Resource Utilization (0.67)	Physician Overload Time	0.40	0.268
	Physician Idle Time	0.30	0.201
	Nurse Overload Time	0.20	0.134
	Nurse Idle Time	0.10	0.067
Patient Satisfaction (0.33)	Patient Waiting Time	0.33	0.109
	No. of Unscheduled Patients	0.67	0.221

Scenario 5

Here, the Borda method is used to determine the weights assuming the same preference structure considered in Scenario 2. Tables 6.26 - 6.28 illustrate the DM's rankings and the corresponding weights for the criteria and sub-criteria. Table 6.29 shows the final weights of the performance measures obtained using the Borda method for Scenario 5.

Table 6.26: Criteria weights using the Borda method for Scenario 5

Criteria	**DM Ranking**	**Points**	**Weights**
Resource Utilization	2	1	0.33
Patient Satisfaction	1	2	0.67

Table 6.27: DM ranking and sub-criteria weights for resource utilization metrics using the Borda method for Scenario 5

Resource Utilization Metrics	**DM Ranking**	**Ranking Points**	**Weights**
Physician Overload Time	1	4	0.40
Physician Idle Time	3	2	0.20
Nurse Overload Time	2	3	0.30
Nurse Idle Time	4	1	0.10

Table 6.28: DM ranking and sub-criteria weights for patient satisfaction metrics using the Borda method for Scenario 5

Patient Satisfaction Metrics	**DM Ranking**	**Ranking Points**	**Weights**
Patient Waiting Time	1	2	0.67
No. of Unscheduled Patients	2	1	0.33

Table 6.29: Final weights of the performance measures obtained using the Borda Method for for Scenario 5

Criteria (Criteria Weight)	**Sub-Criteria**	**Sub-Criteria Weight**	**Final Weight**
Resource Utilization (0.33)	Physician Overload Time	0.40	0.132
	Physician Idle Time	0.20	0.066
	Nurse Overload Time	0.30	0.099
	Nurse Idle Time	0.10	0.033
Patient Satisfaction (0.67)	Patient Waiting Time	0.67	0.449
	No. of Unscheduled Patients	0.33	0.221

It is important to note that the relative weights obtained by the Rating method and the Borda method may be different (see Table 6.30)

6.7.3.3 Other MCDM Method

Another method for determining criteria and sub-criteria weights is the Analytic Hierarchy Process (AHP). AHP is a comparatively complex method and requires sophisticated software to compute the criteria and sub-criteria weights. It also puts more cognitive burden on the DM. It is discussed in detail in the textbook by Ravindran and Warsing (2013).

6.7.3.4 Ranking of the Scheduling Rules under Different Scenarios

The weights obtained for different hospital settings, using different MCDM methods, are summarized in Table 6.30. Note that scenarios 1,2,3 use the Rating method and 4 and 5 use the Borda method. These weights are used to combine the six performance measures to one metric by computing the weighted sum of the performance measures, called the overall score, which is then used to rank the different scheduling rules.

Table 6.30: Weights of the performance measures under different scenarios

Performance Metrics	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
Physician Idle Time	0.256	0.040	0.125	0.201	0.066
Physician Overload Time	0.288	0.066	0.125	0.268	0.132
Nurse Idle Time	0.096	0.034	0.125	0.067	0.033
Nurse Overload Time	0.160	0.060	0.125	0.134	0.099
Patient Waiting Time	0.050	0.600	0.250	0.109	0.449
No. of Unscheduled Patients	0.150	0.200	0.250	0.221	0.221

Unlike the Total Cost method, the MCDM method uses weighted sum of the criteria values and thus requires the performance measures to be scaled appropriately such that they are of similar magnitude. This has to be done because the performance measures may have different units or magnitudes and without scaling, the performance measures with larger magnitudes will dominate the overall score. In our case, the units for resource idle time, resource spillover time and patient waiting time are in minutes, while the units for the number of denied appointments is patients. In this chapter, the performance measures (criteria values) are scaled using the formula given in Equation 6.2.

$$V_{kr} = \frac{a_{kr}}{H_k} \times 100 \tag{6.2}$$

where

V_{kr} - Scaled value of criterion k for scheduling rule r

a_{kr} - Value of criterion k for scheduling rule r

$H_k = \max_r \{a_{kr}\}$ - Maximum value of criterion k across all scheduling rules

Note that the scaled values (V_{kr}) will always be less than or equal to 100 and lower values of V_{kr} are preferred. The scaled performance measures are then used in Equation 6.3 to compute the overall score for each scheduling rule. Since we are interested in minimizing the performance measures, a lower overall score is preferred.

$$S_r = \sum_{k=1}^{K} W_k \times V_{kr} \tag{6.3}$$

Table 6.31 gives the average overall score and its standard deviation over 5000 call-in sequences and 500 replications for various scheduling rules. It can be observed that the proposed scheduling rules perform better (i.e., lower overall score) than the current practice (Round Robin and Evenly Distributed) for all the scenarios. Also, the Tukey multiple comparison test indicates that the proposed rules have significantly lower overall score compared to the current practice. Among the proposed scheduling rules, EABG + OB1 performs the best (i.e., least overall score) for Scenarios 1, 3 and 4 and EABG + OB2 performs the best for Scenarios 2 and 5. In other words, EABG + OB2 appears to be the best, when patient satisfaction is more important over resource utilization, and EABG + OB1 appears to be the best, when resource utilization is important or both patient satisfaction and resource utilization are equally important.

Table 6.31: Average scores and STD for different scenarios

Scheduling Rules	Overall Score									
	Scenario 1		Scenario 2		Scenario 3		Scenario 4		Scenario 5	
	Avg.	STD	Avg.	STD	Avg.	STD	Avg.	STD	Avg.	STD
Round Robin	79.09	5.23	83.42	5.89	76.00	4.75	76.64	5.19	81.00	5.62
Evenly Distributed	69.50	6.39	60.71	9.00	64.08	5.95	65.11	6.48	61.31	8.15
LRBG + OB1	56.81	3.93	48.57	6.20	59.86	4.25	57.03	3.93	50.52	5.33
LRBG + OB2	57.21	3.79	47.98	5.87	59.99	4.08	57.21	3.77	50.10	5.03
HRBG + OB1	57.70	4.05	49.58	6.28	60.75	4.29	57.82	4.06	51.41	5.44
HRBG + OB2	56.81	3.74	48.42	5.89	60.04	4.07	56.96	3.78	50.39	5.10
EABG + OB1	53.19	3.66	46.33	5.91	56.55	4.11	54.01	3.63	48.44	4.99
EABG + OB2	54.52	3.79	45.22	5.81	57.09	4.14	55.00	3.74	47.80	4.99
BIBG + OB1	54.02	3.79	46.33	6.10	56.96	4.22	54.63	3.75	48.58	5.18
BIBG + OB2	54.04	3.69	45.59	5.64	57.28	3.95	54.71	3.56	47.98	4.82

6.8 Managerial Implications

The key finding of this study to healthcare practitioners is that efficient patient sequencing and appointment overbooking could result in substantial improvement to the clinic/hospital performance measures. The results indicate that integrating the risk-type of the patient with the scheduling rules lead to significantly better operational performance compared to the rules currently used in practice, which assume the same no-show rate for all patients and randomly overbook the patients. All the proposed scheduling rules consistently outperform the current practice, namely, round-robin and evenly-distributed rule, under 95% confidence intervals. The sensitivity analysis suggests that our inferences are valid across a reasonable range of no-show rates and service time CV.

6.8.1 Choice of Sequencing Policy

Four sequencing policies are tested in this chapter - two based on patient risk type (LRBG and HRBG) and two based on patient service time (EABG and BIBG). In most cases, the

performance between LRBG and HRBG sequencing, and the performance between EABG and BIBG sequencing did not differ significantly from each other. However, for all no-show rates and service time CV, it is found that sequencing based on expected service times and overbooking based on the risk type of the patient is always superior compared to just the risk type of the patient.

Different outpatient clinics have different patient types (e.g., follow-up, new patients), procedures (e.g., physician visit, surgical procedure) and appointment types (e.g., pre-booked appointments only, combination of pre-booked and same-day appointments). Therefore, the choice of the sequencing/scheduling rule primarily depends on the clinic characteristics. Scheduling high-risk patients in the beginning (HRBG) is more suitable for a clinic that accepts both same-day and pre-booked appointments. This is due to the fact that patients with same day appointments are always low-risk patients and they are generally not available in the beginning of the clinic session as their call for appointment arrives during that time. On the other hand, scheduling low-risk patients in the beginning may be appropriate for clinics that accepts pre-booked patients and has shorter appointment delay. Further, scheduling based on expected service times (either extended appointments in the beginning or brief and intermediate appointments in the beginning) are more suitable for cluster scheduling, where patients are grouped based on certain characteristics. For example, a clinic may schedule all surgical procedures in the afternoon and all office visits in the morning, and therefore could use the BIBG sequencing rule. On the other hand, if a clinic wants to treat all the new patients, who generally require longer service times, in the beginning, and then see established and follow-up patients later, then EABG sequencing rule is more appropriate.

6.8.2 Choice of Overbooking Policy

Two overbooking policies, namely OB1 and OB2 are proposed and analyzed. For all levels of no-show rates and service time CV, the OB1 policy results in lower physician idle time, while the OB2 policy results in lower physician spillover time and patient waiting time.

However, the overbooking policy does not have a substantial impact on nurse idle time, nurse spillover time, number of patients denied appointments. Therefore, if the outpatient clinic is focused on reducing the overburden on the resources and waiting time of the patients, then it is better to adopt OB2 overbooking policy. On the other hand, if the clinic wants to reduce the number of patients denied appointment or the idle time of the physicians, then OB1 policy is preferred. In addition, it is also observed that OB2 policy resulted in significantly lower total cost for scheduling based on risk type of the patient, while OB1 policy resulted in significantly lower cost for scheduling based on expected patient service time. Therefore, depending on the type of sequencing policy, it is always better to adopt one of the two overbooking policies.

The scheduling rules proposed in this chapter can be implemented in practice because it considers realistic clinic conditions, such as multi-phase, multi-provider environment and variable appointment duration. Further, the proposed rules are easy to implement because the amount of information required to generate the schedule is minimal (i.e., appointment duration and risk type of patient). Thus, the proposed sequential framework of using predictive analytics by integrating data from various sources, and then leveraging them to schedule patients could be utilized for creating better decision support tools that help hospital/clinic administrators to generate effective appointment schedules and best utilize their valuable human resources.

6.9 Conclusions

The scheduling framework proposed in Chapter 5 has three steps - (1) Data collection, (2) predictive analytics to classify the patients based on their no-show risk and (3) use scheduling rules to leverage the information obtained from predictive analytics. In this chapter, the application of proposed framework is illustrated using a case study with real data obtained from Penn State Hershey Medical Center (PSHMC). First, the data collected from PSHMC and the associated clinic's setting are discussed in detail. Using that information, the parameters required for the predictive analytics (Step 2) and the simulation

of proposed scheduling rules (Step 3) are illustrated.

Five different machine learning (ML) algorithms were evaluated and Stacking algorithm was found to be the best based on its AUC value for this case study. Thus, based on the data obtained, Stacking algorithm was recommended for classifying the patients based on their risk of no-shows. Further, all the proposed scheduling rules use the risk of no-show obtained from the Stacking algorithm to schedule patients.

The scheduling rules, both current practice and the proposed rules, are evaluated by two approaches : (1) using relative costs associated with the performance metrics; (2) using MCDM method to get relative weights associated with the performance metrics. The evaluation of the scheduling rules by both the approaches indicate that integrating the risk of patient no-show with patient scheduling leads to significantly better operational performance compared to rules that randomly overbooks patients assuming the same no-show rate. Further, scheduling patients based on expected service times and overbooking a combination of low-risk and high-risk patient in the first available slot consistently result in the least total cost.

The analysis also indicates that there is no single rule that results in the least overall value for all the six performance measures. Therefore, future work may consider the use of interactive multiple criteria decision making approaches to identify the best compromise solution and provide recommendations on the choice of scheduling rules for different clinic settings. The increasing healthcare costs and high demand for outpatient services can be effectively handled by using the proposed framework. Although the results of this research are specific to a healthcare facility, the proposed framework has applicability to other service facilities that experiences significant no-shows and has appointment based arrivals, such as beauty salons.

Chapter 7

Conclusions and Future Work

The demand for outpatient services is rising and hospitals are expected to receive 65% of their revenue from outpatient care. However, the capacity of the outpatient clinics are limited by its operating hours and physician availability. Outpatient clinics use an appointment system to handle the increasing demand with existing capacity. Traditional scheduling methods, such as pre-booking patients weeks in advance, may not be suitable in today's healthcare setting as patient's expectations (in terms of clinic accessibility and waiting time) continues to increase. Moreover, with the rise in the adoption of Electronic Health Records (EHR) at the hospitals, more patient data will be available and can be used to design effective appointment systems.

Designing an appointment system involves multiple decisions, such as determining the appointment types (same-day and pre-scheduled), patient sequence, and appointment time. Further, various factors, such as patient flow, demand uncertainty and patient no-shows must be considered to develop an effective design. Inefficiencies in the appointment system design and patient no-shows (patients who do not arrive for scheduled appointments) cost the U.S. healthcare system more than $150 billion a year (Toland, 2013). In addition, they also reduce provider productivity and timely access to care. This dissertation focused on the design of data-driven appointment systems for outpatient hospitals and clinics such that the needs of the patient as well as the hospitals/clinics are met.

7.1 Data Source

The data used in this dissertation have been obtained from the Penn State Hershey Medical Center (PSHMC) located in central Pennsylvania. PSHMC is part of the Pennsylvania State University and the Penn State Health Care Partners, clinically integrated network of regional care providers. It has over 1,100 physicians. In 2016, PSHMC treated over million outpatients. The data used in this dissertation are approved under PSHMC institutional review board (IRB) *STUDY00004553: Design and Analysis of Appointment Systems for Outpatient Clinics*. The data included all patient visits from September 2014 to August 2016 with detailed action description. The action description allowed us to individually track each patient visit. For example, for a particular patient the action description indicates when a patient called for an appointment, when the patient scheduled an appointment, whether the patient modified, canceled or showed-up for the appointment, and what time the patient visited the clinic on the day of their appointment.

7.2 Theoretical Contributions

This dissertation has several contributions to the theory of designing appointment systems for outpatient clinics. Based on the review of literature in Chapter 2, it was evident that almost all the previous work on appointment system design considered a simplified clinic setting with single phase and single provider. However, in practice, most outpatient departments have multi-phase (e.g., pre-screening, visit nurse, visit doctor, check-out) multi-provider setting. In addition, most of the previous work rarely considered patient's provider preference, patient availability, uncertainty in patient demand, patient no-shows and service time. In practice, a patient may have a specific provider preference and may be available only during certain times of the day for a hospital visit. A detailed analysis in Chapter 3 indicated that ignoring the multi-phase nature of patient flow, patient's provider preference and patient's availability lead to unmet demand, patient dissatisfaction and inefficient resource utilization. Moreover, appointment systems based on a simplified clinic

setting are not applicable to a multi-phase, multi-provider setting. In this dissertation, the appointment system is designed by considering realistic clinic settings (i.e., multi-phase, multi-provider, patient's provider preference, patient's availability, demand and service time uncertainty). Thus, it bridges the gap between theory and practice. In particular, two different appointment systems have been proposed and studied.

First, the design of a hybrid appointment system (a combination of open access and pre-booking appointment type) was studied by integrating the multi-phase multi-provider setting, patient preference and uncertainties in arrival and service times. To achieve the best outcomes in a hybrid appointment system, it is essential to determine the number of slots reserved for each appointment type. Overestimating the total number open access appointments leads to increased resource idle time, while underestimating it increases appointment delay. The objective of the research work was to determine the best schedule configuration for a realistic clinic setting that achieves a balance between resource utilization and patient satisfaction.

Second, sequential scheduling rules (sequencing and overbooking decisions) for designing the appointment systems have been developed and studied. Most of the existing rules are based on the experience of the hospital administrator or average values of clinic characteristics, such as service times and no-show rates. None of the appointment rules in the literature is based on individual patient characteristics, such as risk of no-show. Therefore, eight data-driven scheduling rules, which use patient specific no-show value, have been proposed and analyzed for their effectiveness compared to current practice.

Unlike most of the previous work, all the parameters (e.g., patient calls for appointment, service time) associated with the system design are data-driven. In other words, for all the parameters, realistic data was used to fit theoretical distributions to study the clinic settings observed in practice as well as clinic settings expected to occur in the future.

7.3 Methodological Contributions

The dissertation also contributes to methodological advances. For designing the hybrid appointment systems, a new methodology is proposed by using a combination of optimization, Monte-Carlo scenario analysis, and heuristics. A deterministic model is proposed to determine the optimal percentage of appointments reserved for pre-booking and open access, and a scenario-based Monte Carlo approach is used to account for uncertainty. Finally, the best configuration for the hybrid appointment system is determined using two proposed heuristics.

For determining the sequential scheduling rules associated with the appointment system, a three step scheduling framework that uses a combination of predictive analytics and simulation was proposed. First, the patient-related data from various sources are extracted. Then, the best machine learning algorithm is selected to predict the individual risk of patient no-show based on multiple patient characteristics and environmental conditions. Finally, eight scheduling rules have been proposed, which use the risk of no-show for sequencing and/or overbooking decisions. The proposed framework is generic and therefore, has applicability in other non-medical facilities that experience no-shows and have appointment based arrivals, such as beauty salons.

7.4 Practical Implications

There are several key findings in this dissertation for healthcare practitioners. When designing hybrid appointment systems, it is best to overbook patients in the beginning of the clinic session to compensate for no-shows and ensure steady patient flow for physicians to begin the day. Also, when double-booking a slot, at least one of the patient should be high risk of no-show. The same day appointments must be later in the clinic session and must not be double-booked.

As far as scheduling rules are considered, efficient patient sequencing and appointment overbooking could result in substantial improvements to the clinic/hospital performance

measures, compared to the first come first appointment basis and/or random scheduling. In particular, using Electronic Health Records to identify patient specific no-show rate and then integrating the risk of patient no-show with sequencing and/or overbooking decisions lead to significantly better operational performance. Further, the choice of sequencing and overbooking decisions depends on the clinic setting.

Overall, the key lesson for practitioners is that the EHR data can be leveraged to design appointment systems and improve health outcomes. In addition, application of engineering principles could revitalize the healthcare system.

7.5 Scope for Future Research

This research work can be extended with respect to theory and methodology. In this dissertation, the hybrid appointment system included a combination of same day and pre-booked appointments. For future research, the presence of “walk-ins" can be included to design the appointment system. A hybrid appointment system with walk-ins can have a tremendous impact in reducing emergency room waiting time, because about 71% of the emergency department visits are avoidable or unnecessary. Therefore, these patients can be given same-day appointments or pre-booked appointments within the next 72 hours, and the real emergency cases can be treated as walk-in appointments. Thus, a hybrid system with walk-in appointments is applicable for emergency rooms. Alternatively, a hybrid appointment systems that incorporate virtual appointments can also be studied. Recently, there is growing consumer interest for virtual appointments, where the patients are not required to visit the clinic for service. Moreover, virtual appointments reduce patient visits to the clinic, long waiting times, and waiting room congestion. Therefore, incorporating virtual appointments provides greater flexibility to the appointment system design.

With respect to the methodology extension, a simulation based meta-heuristic approach can be developed to obtain the best schedule configuration of the hybrid appointment system with less computational effort (in terms of time and scenarios) compared to the mathematical model. Also, goal programming and interactive multi-criteria approaches,

where the decision makers are involved throughout the solution process, can be developed as there are multiple conflicting performance measures to evaluate an appointment system design.

Appendix A

Appointment Time Calculation for different Appointment Rules

Rule	Appointment Time	
IBFI	$T(O_{m,d,s}) = 0$	$\forall m = 1, d, s = 1$
	$T(O_{m,d,s}) = T(O_{m-1,d,s}) + \mu$	$\forall m > 1, d, s > 1$
2ATBEG	$T(O_{m,d,s}) = 0$	$\forall m \leq 2, d, s = 1$
	$T(O_{m,d,s}) = T(O_{m-1,d,s}) + \mu$	$\forall m > 2, d, 1 < s \leq S$
LVBEG	$T(O_{m,d,s}) = (s-1)\mu$	$\forall m \in LV, d, s \leq S_1$
	$T(O_{m,d,s}) = S\mu - (s - S_1)\mu$	$\forall m \in HV, d, S - S_1 < s \leq S$
HVBEG	$T(O_{m,d,s}) = (s-1)\mu$	$\forall m \in HV, d, s \leq S_2$
	$T(O_{m,d,s}) = S\mu - (s - S_2)\mu$	$\forall m \in LV, d, S - S_2 < s \leq S$
OFFSET	$T(O_{m,d,s}) = (m-1)\mu - \beta_1(k-m)\sigma$	$\forall m \leq k, d, s \leq k$
	$T(O_{m,d,s}) = (m-1)\mu + \beta_2(m-k)\sigma$	$\forall m > k, d, k < s \leq S$
DOME	$T(O_{m,d,s}) = (m-1)\mu - \beta_1(k_1-m)\sigma$	$\forall m \leq k_1, d, s \leq k_1$
	$T(O_{m,d,s}) = (m-1)\mu + \beta_2(m-k_1)\sigma$	$\forall d, k_1 < m, s < k_2 \; \forall m \geq k_2, d, k_2 \leq s \leq S$
	$T(O_{m,d,s}) = (m-1)\mu - \beta_3(m-k_2)\sigma$	

where

m	Denotes the index of the m^{th} patient in the schedule
LV	Set of low-variance patients
HV	Set of high-variance patients
μ	Average service time (in minutes)
σ	Standard deviation of service time (in minutes)
S_1	Total number of slots for low variance patients
S_2	Total number of slots for high variance patients
k	Slots up to which patients are scheduled to arrive earlier in OFFSET rule
k_1	Slots up to which patients are scheduled to arrive earlier in DOME rule
k_2	Slots up to which patients are scheduled to arrive later in DOME rule
$O_{m,d,s}$	Denotes the patient index of m^{th} patient on day d in slot s
$T(O_{m,d,s})$	Appointment time of m^{th} patient on day d in slot s

Appendix B

Pseudo-code for Algorithm 1

Step 1: Select the appointment rule (U) to schedule all the calling patients in the scheduling horizon.

Step 2: Initialize p=1, d=1, s=1 and $A_{p,d,s} = 0$.

Step 3:: Determine if patient p can be scheduled in slot s by evaluating the two conditions described below. If one of the two conditions is satisfied, then go to Step 4. Else, go to Step 8.

- If U = 2ATBEG rule and $s = 1$, then the number of patients scheduled in slot s must be less than equal to 1.
- If U = 2ATBEG rule and $s > 1$ or if U = {IBFI, LVBEG, HVBEG, OFFSET, DOME} and $s \geq 1$, then the number of patients scheduled in slot s must be 0.

Step 4: For LVBEG and HVBEG rule, classify the patient as either high variance (β_p = 1) or low variance ($\beta_p = 0$) based on his service time variation. For all other appointment rules, go to Step 6.

Step 5a: For LVBEG rule, if patient p is low-variance patient and if slot s is greater than S_1, then go to Step 6. Else, go to Step 8.

Step 5b: For LVBEG rule, if patient p is high-variance patient and if slot s is greater than $S - S_1$ for high variance patient, then go to Step 6. Else, go to Step 8.

Step 5c: For HVBEG rule, if patient p is low-variance patient and if slot s is greater than $S - S_2$, then go to Step 6. Else, go to Step 8.

Step 5d: For HVBEG rule, if patient p is high-variance patient if slot s is greater than S_2, the go to Step 6. Else, go to Step 8.

Step 6: Obtain the availability of patient p on day d in slot s ($\alpha_{p,d,s}$).

Step 7: If patient p is available on day d in slot s, then schedule patient p in slot s of day d, set $A_{p,d,s}$ = 1, and go to Step 11. Else, go to Step 8.

Step 8: Set $s = s + 1$. If s $\leq$ S, then go to Step 3. Else, go to Step 10.

Step 9: Set $d = d + 1$, $s = 1$. If $d \leq D$, then go to Step 3. Else, go to Step 12.

Step 10: Patient p cannot be scheduled. Set $A_{p,d,s}$ = 0, and go to Step 11.

Step 11: Set $p = p + 1$. If $p \leq P$, then go to step 3. Else, go to Step 12.

Step 12: STOP. The schedule construction is complete.

Appendix C

Pseudo-code for Algorithm 2

Step 1: Obtain the index of the patient (p) scheduled on all days across all slots from the fully constructed schedule.

Step 2: Initialize $i = 1$, $d = 1$, $s = 1$, $h = 1$, $WT_{p,i} = 0$, $OT_{h,d,i} = 0$, $IT_{h,d,s,i} = 0$

Step 3: Generate the service time of patient p in phase h and replication i ($ST_{p,h,i}$) and the no-show value associated with patient p for replication i ($\eta_{p,i}$).

Step 4: Determine the actual start time, $AST_{p,h,i}$, and actual completion time, $ACT_{p,h,i}$, of patient p in phase h and replication i.

Step 5: Set $h = h+1$. If $h \leq H$, then go to step 3. Else, go to step 6.

Step 6: Set $s = s+1$ and $h = 1$. If $s \leq S$, then go to step 2. Else, go to step 7.

Step 7: Set $d = d+1$, $s = 1$ and $h = 1$. If $d \leq D$, then go to step 2. Else, go to step 8.

Step 8: For replication i, calculate the waiting time of patient p ($WT_{p,i}$), idle time of the resource in phase h on day d in slot s ($IT_{h,d,s,i}$), and overtime of resource in phase h on day d ($OT_{h,d,i}$).

Step 9: Set $i = i+1$, $d = 1$, $s = 1$ and $h = 1$. If $i \leq I$, then go to step 4. Else, go to Step 10.

Step 10: Compute the average waiting time ($\overline{WT}$), average idle time ($\overline{IT}$), and average overtime ($\overline{OT}$) across all replications, days, slots, phases and patients.

Step 11: STOP. Schedule evaluation is complete.

Bibliography

[1] Ahluwalia, S., & Offredy, M. (2005). A qualitative study of the impact of the implementation of advanced access in primary healthcare on the working lives of general practice staff. *BMC family practice*, 6(1), 1.

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

[3] Barnes, C. D., Quiason, J. L., Benson, C., & McGuiness, D. (1997, December). Success stories in simulation in health care. In *Proceedings of the 29th conference on Winter simulation*, IEEE Computer Society, pp. 1280-1285.

[4] Barron, W. M. (1980). Failed appointments. Who misses them, why they are missed, and what can be done. *Primary care*, 7(4), 563-574.

[5] Birge, J. R., & Louveaux, F. (2011). Introduction to stochastic programming. *Springer Science & Business Media*.

[6] Boelke, C., Boushon, B., & Isensee, S. (1999). Achieving open access: the road to improved service & satisfaction. *Medical group management journal/MGMA*, 47(5), 58-62.

[7] Brahimi, M., & Worthington, D. J. (1991). Queueing models for out-patient appointment systems–a case study. *Journal of the Operational Research Society*, 733-746.

[8] Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4), 519-549.

[9] Cayirli, T., Veral, E., & Rosen, H. (2006). Designing appointment scheduling systems for ambulatory care services. *Health care management science*,9(1), 47-58.

[10] Cayirli, T., Yang, K. K., & Quek, S. A. (2012). A Universal Appointment Rule in the Presence of No-Shows and Walk-Ins. *Production and Operations Management*, 21(4), 682-697.

[11] Cayirli, T., & Gunes, E. D. (2013). Outpatient appointment scheduling in presence of seasonal walk-ins. *Journal of the Operational Research Society*,65(4), 512-531.

[12] Chakraborty, S., Muthuraman, K., & Lawley, M. (2010). Sequential clinical scheduling with patient no-shows and general service time distributions. *IIE Transactions*, 42(5), 354-366.

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

[14] Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., DeLaurentis, P. C., ... & Sands, L. (2010). Using no-show modeling to improve clinic performance. *Health Informatics Journal*, 16(4), 246-259.

[15] Dharmapala, P.S. (2009). Adding value in healthcare service by improving operational efficiency using data envelopment analysis, *International Journal of Operational Research*, Vol. 5, No. 1, 73-88.

[16] Dreiher, J., Froimovici, M., Bibi, Y., Vardy, D. A., Cicurel, A., & Cohen, A. D. (2008). Nonattendance in obstetrics and gynecology patients. *Gynecologic and obstetric investigation*, 66(1), 40-43.

[17] El-Sharo, M., Zheng, B., Yoon, S. W., & Khasawneh, M. T. (2015). An overbooking scheduling model for outpatient appointments in a multi-provider clinic.*Operations Research for Health Care*, 6, 1-10.

[18] Erdogan, S. A., & Denton, B. (2013). Dynamic appointment scheduling of a stochastic server with uncertain demand. *INFORMS Journal on Computing*, 25(1), 116-132.

[19] Erdogan, S. A., Gose, A., & Denton, B. T. (2015). Online appointment sequencing and scheduling. *IIE Transactions*, 47(11), 1267-1286.

[20] Fetter, R. B., & Thompson, J. D. (1966). Patients' waiting time and doctors' idle time in the outpatient setting. *Health Services Research*, 1(1), 66.

[21] Field, J. (1980). Problems of urgent consultations within an appointment system. *JR Coll Gen Pract*, 30(212), 173-177.

[22] Fries, B. E., & Marathe, V. P. (1981). Determination of optimal variable-sized multiple-block appointment systems. *Operations Research*, 29(2), 324-345.

[23] Geraghty, M., Glynn, F., Amin, M., & Kinsella, J. (2008). Patient mobile telephone "text" reminder: a novel way to reduce non-attendance at the ENT out-patient clinic. *The Journal of Laryngology & Otology*, 122(03), 296-298.

[24] Glowacka, K. J., Henry, R. M., & May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society*, 60(8), 1056-1068.

[25] Gul, S., Denton, B. T., Fowler, J. W., & Huschka, T. (2011). Bi-Criteria Scheduling of Surgical Services for an Outpatient Procedure Center.*Production and Operations management*, 20(3), 406-417.

[26] Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9), 800-819.

[27] Hahn-Goldberg, S., Carter, M. W., & Beck, J. C. (2012). Dynamic template scheduling to address uncertainty in complex scheduling problems: a case study on chemotherapy outpatient scheduling. In *Society for Health Systems Conference*, Las Vegas, NV.

[28] Herriott, S. (1999). Reducing delays and waiting times with open-office scheduling. *Family practice management*, 6, 38-43.

[29] Ho, C. J., & Lau, H. S. (1992). Minimizing total cost in scheduling outpatient appointments. *Management science*, 38(12), 1750-1764.

[30] Ho, C. J., Lau, H. S., & Li, J. (1995). Introducing variable-interval appointment scheduling rules in service systems. *International Journal of Operations & Production Management*, 15(6), 59-68.

[31] Ho, C. J., & Lau, H. S. (1999). Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operational Research*, 112(3), 542-553.

[32] Izard, T. (2005). Managing the habitual no-show patient. *Family practice management*, 12(2), 65-66.

[33] Kaandorp, G. C., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3), 217-229.

[34] Kennedy, J. G., & Hsu, J. T. (2003). Implementation of an open access scheduling system in a residency training program. *FAMILY MEDICINE*, 35(9), 666-670.

[35] Klassen, K. J., & Rohleder, T. R. (1996). Scheduling outpatient appointments in a dynamic environment. *Journal of operations Management*, 14(2), 83-101.

[36] Kopach, R., DeLaurentis, P. C., Lawley, M., Muthuraman, K., Ozsen, L., Rardin, R., Wan, H., Intrevado, P., Qu, X., & Willis, D. (2007). Effects of clinical characteristics on successful open access scheduling. *Health care management science*, 10(2), 111-124.

[37] Kuhn, M. (2008). Caret package. Journal of Statistical Software, 28(5), 1-26.

[38] LaGanga, L. R., & Lawrence, S. R. (2012). Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management*, 21(5), 874-888.

[39] Lee, S., & Yih, Y. (2010). Analysis of an open access scheduling system in outpatient clinics: a simulation study. *Simulation*, 86(8-9), 503-518.

[40] Lindh, W., Pooler, M., Tamparo, C., & Dahl, B. (2009). Delmar's administrative medical assisting. *Cengage Learning*.

[41] Lindley, D. V. (1952, April). The theory of queues with a single server. In *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 48, No. 02, pp. 277-289). Cambridge University Press.

[42] Liu, N., Ziya, S., & Kulkarni, V. G. (2010). Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management*, 12(2), 347-364.

[43] Merritt Hawkins & Associates (2009) Survey of Physician Appointment Wait Times, pp.1-17, Irving, TX [online] `http://www.merritthawkins.com/pdf/mha2009waittimesurvey.pdf` (accessed 15 May 2015).

[44] Mogha, S.K., Yadav, S.P. and Singh, S.P. (2014) Estimating technical and scale efficiencies of private hospitals using a non-parametric approach: case of India. *International Journal of Operational Research*, Vol. 20, No. 1, 21-40.

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

[46] Murray, M., & Berwick, D. M. (2003). Advanced access: reducing waiting and delays in primary care. *JAMA*, 289(8), 1035-1040.

[47] Muthuraman, K., & Lawley, M. (2008). A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*, 40(9), 820-837.

[48] Naidu, A. (2009). Factors affecting patient satisfaction and healthcare quality. *International journal of health care quality assurance*, 22(4), 366-381.

[49] National Health Expenditure Projections (2014) National Health Expenditure Projections 2013-2023, *Centers for Medicare and Medicaid Services*, 25 May, Web. 21 January 2015.

[50] National Health Expenditure Projections (2015) National Health Expenditure Projections 2014-2024, *Centers for Medicare and Medicaid Services*, 25 May, Web. 15 October 2015.

[51] O'Hare, C. D., & Corlett, J. (2004). The outcomes of open-access scheduling. *Family practice management*, 11(2), 35-38.

[52] Oaklander, M. (2015). Doctors on Life Support. *TIME Magazine (U.S. edition)*, Vol 186, No.9.

[53] Patrick, J. (2012). A Markov decision model for determining optimal outpatient scheduling. *Health care management science*, 15(2), 91-102.

[54] Peng, Y., Qu, X., & Shi, J. (2014). A hybrid simulation and genetic algorithm approach to determine the optimal scheduling templates for open access clinics admitting walk-in patients. *Computers & Industrial Engineering*, 72, 282-296.

[55] 57. Pinedo, M. (2005). Planning and scheduling in manufacturing and services (Vol. 24). *New York: Springer*.

[56] Qu, X., & Shi, J. (2011). Modeling the effect of patient choice on the performance of open access scheduling. *International Journal of Production Economics*, 129(2), 314-327.

[57] Qu, X., Rardin, R. L., & Williams, J. A. S. (2011). Single versus hybrid time horizons for open access scheduling. *Computers & Industrial Engineering*,60(1), 56-65.

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

[59] Rising, E. J., Baron, R., & Averill, B. (1973). A systems analysis of a university-health-service outpatient clinic. *Operations Research*, 21(5), 1030-1047.

[60] Robinson, L. W., & Chen, R. R. (2010). A comparison of traditional and open-access policies for appointment scheduling. *Manufacturing & Service Operations Management*, 12(2), 330-346.

[61] Rohleder, T. R., & Klassen, K. J. (2000). Using client-variance information to improve dynamic appointment scheduling performance. *Omega*, 28(3), 293-302.

[62] Saremi, A., Jula, P., ElMekkawy, T., & Wang, G. G. (2013). Appointment scheduling of outpatient surgical services in a multistage operating room department. *International Journal of Production Economics*, 141(2), 646-658.

[63] Shonick, W., & Klein, B. W. (1977). An approach to reducing the adverse effects of broken appointments in primary care systems: development of a decision rule based on estimated conditional probabilities. *Medical Care*, 419-429.

[64] Singer, I. A., & Regenstein, M. (2003). Advanced Access: Ambulatory Care Redesign and the Nation's Safety Net. National Association of Public Hospitals and Health Systems.

[65] Soriano, A. (1966). Comparison of two scheduling systems. *Operations Research*, 14(3), 388-397.

[66] Srinivas, S. (2015). Mathematical Programming Approach for Patient Scheduling in a Hybrid Appointment System. *M.Eng Paper, The Pennsylvania State University*.

[67] Taylor, B. (1984). Patient use of a mixed appointment system in an urban practice. *BMJ*, 289(6454), 1277-1278.

[68] Toland, B. (2013). No-shows cost health care system billions. Retrieved April 28, 2017, from http://www.post-gazette.com/business/businessnews/2013/02/24/No-shows-cost-health-care-system-billions/stories/201302240381/

[69] Virji, A. (1990). A study of patients attending without appointments in an urban general practice. *BMJ*, 301(6742), 22-26.

[70] Yan, C., Tang, J., Jiang, B., & Fung, R. Y. (2015). Sequential appointment scheduling considering patient choice and service fairness. *International Journal of Production Research*, 53(24), 7376-7395.

Vita

Sharan Srinivas

Sharan Srinivas was born on July 15, 1990 in Chennai, Tamil Nadu, India. Raised by Srinivasan and Poongothai, he attended schools in Chennai. In May of 2011, Sharan graduated from College of Engineering, Guindy, Anna University, where he earned his Bachelor of Engineering degree in industrial engineering. In August of 2011, Sharan came to the United States of America as a Graduate Research Associate to conduct research and graduate studies in the Department of Systems Science and Industrial Engineering at Binghamton University, State University of New York. While pursuing his MS degree, he simultaneously worked with New York State University Police, International Student and Scholar Services, and Howard Hughes Medical Institute on various projects. Also, he completed his Lean Six Sigma Black Belt certification during this time.

In August of 2013, Sharan earned his Master of Science degree in industrial and systems engineering from Binghamton University and joined the Department of Industrial and Manufacturing Engineering at the Pennsylvania State University (Penn State) to pursue his doctorate degree. During this time, he also earned his Master of Engineering degree in industrial engineering and operations research from Penn State. At Penn State, his research interests were in the areas of big data analytics, healthcare delivery systems, operations research and industrial engineering. In particular, he worked on designing data-driven healthcare appointment systems for patient scheduling in outpatient clinics. He collaborated with Hershey Medical Center in Pennsylvania to obtain data for his research and test the proposed approaches. In addition, he has been involved in various collaborative research projects that were sponsored by private and public institutions. He developed models and decision support systems to optimize the workforce of the US Army Medical Department, predict power outage occurrences of an electric utility company, identify non-performing suppliers of Volvo, and optimize student workforce assignment of Holy Family Academy. Apart from conducting research, he also served as a teaching assistant for over two years and as an instructor for one semester for a senior level undergraduate course at Penn State.

Sharan has published many peer-reviewed articles in journals such as *Interfaces, International Journal of Operational Research, International Journal of Logistics Systems and Management*. He has co-authored two chapters in the book titled *Big Data Analytics using Multi-Criteria Decision Making Models* to be published by CRC prss in June 2017. He is a recipient of the INFORMS Koopman prize, Penn State Doctoral Fellowship and Service Enterprise Engineering Fellowship. His research interests include healthcare operations management, service systems engineering, supply chain management, application of data analytics and multiple criteria optimization to manufacturing and service systems. He is an active member of INFORMS and IISE professional societies. He has served numerous times as Session Chair for the INFORMS Annual Meeting and IISE Annual Conference.

Upon completing his doctoral degree, Sharan will join the University of Missouri, Columbia as an Assistant Professor with a joint appointment in the Department of Industrial and Manufacturing Systems Engineering and Department of Marketing.