Multi-objective Operating Room Planning and Scheduling

by

Qing Li

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

ARIZONA STATE UNIVERSITY

December 2010

Multi-objective Operating Room Planning and Scheduling

by

Qing Li

has been approved

September 2010

Graduate Supervisory Committee:

John Fowler, Co-Chair
Srimathy Mohan, Co-Chair
Mohan Gopalakrishnan
Ronald Askin
Teresa Wu

ACCEPTED BY THE GRADUATE COLLEGE

ABSTRACT

Surgery is one of the most important functions in a hospital with respect to operational cost, patient flow, and resource utilization. Planning and scheduling the Operating Room (OR) is important for hospitals to improve efficiency and achieve high quality of service. At the same time, it is a complex task due to the conflicting objectives and the uncertain nature of surgeries. In this dissertation, three different methodologies are developed to address OR planning and scheduling problem. First, a simulation-based framework is constructed to analyze the factors that affect the utilization of a catheterization lab and provide decision support for improving the efficiency of operations in a hospital with different priorities of patients. Both operational costs and patient satisfaction metrics are considered. Detailed parametric analysis is performed to provide generic recommendations. Overall it is found the 75^{th} percentile of process duration is always on the efficient frontier and is a good compromise of both objectives. Next, the general OR planning and scheduling problem is formulated with a mixed integer program. The objectives include reducing staff overtime, OR idle time and patient waiting time, as well as satisfying surgeon preferences and regulating patient flow from OR to the Post Anesthesia Care Unit (PACU). Exact solutions are obtained using real data. Heuristics and a random keys genetic algorithm (RKGA) are used in the scheduling phase and compared with the optimal solutions. Interacting effects between planning and scheduling are also investigated. Lastly, a multi-objective simulation optimization approach is developed, which relaxes the deterministic assumption in the second study by

integrating an optimization module of a RKGA implementation of the Non-dominated Sorting Genetic Algorithm II (NSGA-II) to search for Pareto optimal solutions, and a simulation module to evaluate the performance of a given schedule. It is experimentally shown to be an effective technique for finding Pareto optimal solutions.

I dedicate this work to my advisors, Dr. John Fowler and Dr. Srimathy Mohan, my mom and dad, and all my friends who supported me in any respect during the completion of this work.

ACKNOWLEDGMENTS

I have many people to thank for their help and support. First of all, I would like to express my sincere gratitude and respect to my advisors, Dr. John Fowler and Dr. Srimathy Mohan. This work would not have been possible without your consistent support and invaluable guidance along the way. Thank you for providing such excellent professional and personal models for me to carry to the next path in my career and life.

I greatly appreciate my committee members, Dr. Mohan Gopalakrishnan, Dr. Ronald Askin and Dr. Teresa Wu. Their guidance, suggestions and encouragement helped me finish the dissertation smoothly.

Additionally, I am heartily thankful to all my friends in the Industrial Engineering department. Without them, my experience at Arizona State University would not have been full of happiness and fruitfulness. Just to name a few: Shanshan Wang, Junzilan Cheng, Liangjie Xue, Erika Murguia, Mengying Fu, Brinton MacMillan, Zhuoyang Zhou, Wandaliz Torres Garcia, Ozgur Araz, Serhat Gul, James Broyles.

Last but not least, I would like to thank my family and friends. Thank my mom and dad for their unconditional love, faith, and support. I am blessed to have a great family. Thank Bonnie Lerberg, Na Wang and Tingting Ma for their encouragement, company and love.

I am grateful to have met so many wonderful people since I came to Arizona. Without all of your support, this process would have been a much more difficult one. I love you all sincerely.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

1. Introduction and motivation

Between 1999 and 2007 in the United States, healthcare consumed 35.7% of the real increase per capita income, and the share of (Gross Domestic Product) GDP devoted to healthcare rose from 13.7% to 16.2% (Chernew et al., 2009). In 2007, total health care spending in the United States reached $2.3 trillion (Erdogan & Denton, 2009). A report forecast that the healthcare cost could rise to 34% of GDP in three decades unless something was done to overhaul the industry ("Moving up", 2009). In this context, hospitals face an increasing pressure for high quality care and cost effectiveness. As one of the key hospital resources, OR is accounting for 40% of a hospital's resource costs (Marcario et al., 1995). The activities in the OR also have a dramatic impact on many other activities within a hospital. Consequently, the OR department should be continuously enhance quality and lower cost.

Recent studies have shown that the most costs of surgical procedures consist of personnel, infrastructure, equipment, logistics and administrative support, not of materials expense (Roland et al., 2006). The constraint environment has driven the need for efficient resource usage. At the same time, OR planning and scheduling is challenging. Firstly, multiple stakeholders with conflicting interests are involved (Glouberman & Mintzberg, 2001) such as surgeons of various specialties, OR personnel, and patients. Secondly, OR surgical scheduling is complicated by the uncertainty regarding the occurrence

and duration of surgeries. The arrival of non-elective patients may disrupt the planned scheduling throughout the day. The inherent variation and unpredicted nature of the surgeries also causes modifications to fixed schedules. Lastly, the OR department is facing conflicting performance criteria: high planned utilization may lead to excessive patient waiting, while allocating more time to a surgery to decrease the waiting could give rise to staff overtime. This problem has thus attracted the attention of many researchers (Dexter & Traub, 2002; Cardoen & Demeulemeester, 2007; Hans et al., 2008; Jebali et al., 2006).

2. Organization of the dissertation

The objective of this dissertation is focused on capacity planning and scheduling to support managerial decision making in hospitals. We construct models of operating room planning and scheduling to improve the efficiency and quality of service. The reminder of this dissertation is organized as follows.

In Chapter 2, a simulation model is developed to evaluate the performance of the existing approach and compare alternative policies at the catheterization lab, a type of operating room, at a local hospital in Arizona. In this chapter we focus on the day-to-day patient scheduling problem and try to compromise to conflicting objectives with considerations of three types of patients with different priorities. The factors that we evaluated are the size of time block assigned to each procedure, procedure duration, arrival of emergency patients, as well as variation in demand. We consider both operational costs and patient satisfaction metrics, such that decision makers can trade-off between the two metrics. Detailed parametric analysis is performed to develop generic recommendations.

In Chapter 3, the general OR planning and scheduling problem is decomposed into two phases, which are cyclic block scheduling phase and day-to-day patient scheduling phase. It is formulated in mixed integer programming and then solved with CPLEX. The objectives of the model include reducing staff overtime, idle time and patient waiting time, as well as satisfying the surgeons' preference and minimizing the number of beds used in the PACU. Heuristics and RKGA are used in the daily patient sequencing and compared with the optimal solutions from the mathematical model. We will also investigate the necessity of interacting both phases.

Chapter 4 applies simulation optimization methodology in the OR scheduling problem. We develop a multi-objective simulation optimization approach, which integrates an optimization module of RKGA and NSGA-II to guide the search of Pareto optimal solutions, and a simulation module to evaluate the performance of a given schedule. We examine the effectiveness of the approach using real surgical data and compare with alternative approaches. Some managerial questions in OR scheduling are also analyzed. The dissertation concludes with final remarks in Chapter 5.

CHAPTER 2

IMPROVING THE EFFICIENCY OF CATHETERIZATION LABORATORIES USING SIMULATION

1. Introduction

Cardiac catheterization is a diagnostic procedure that comprehensively examines the functioning of the heart and its blood vessels and is usually performed diagnostically, prior to heart surgery. As the size of the population suffering from cardiac problems increases, the number of catheterization procedures performed is growing rapidly. From 1979 to 2002, the number of cardiac catheterizations in the USA increased by 390% and in Europe from 1992 to 1999 by 112% (Katzberg & Haller, 2006), making catheterizations one of the fastest-growing clinical services.

Catheterization laboratories (cath labs) have high fixed and operating costs associated with facilities and staff salaries, and hence, using the lab's time as efficiently as possible becomes crucial to hospital managers and helps them control costs associated with cath labs. Uncertainties in patient arrival and service times along with the varying degree of patient urgency complicate the process of efficient planning, leading to overall poor capacity utilization of resources, recurring staff overtime and excessive patient waiting time (Gupta & Denton, 2008). Appointment systems that assign a specific time window for a case, referred to as *block scheduling*, improve utilization of resources and also allow physicians to know case start times well in advance (Ozcan, 2005). However, these systems generally do not provide the ability for analyzing the impact of

system uncertainties and critical variables such as the block-size, as well as the impact of dynamic rescheduling of delayed patients on lab idle time, staff overtime and patient waiting time.

For cath labs, two factors are major contributors to excessive staff overtime and patient waiting time. *First*, the inherent variation and unpredictable nature of these procedures can cause disruptions or modifications to fixed schedules. The service times are diagnosis-dependent and can vary substantially across patients and surgeons (Gupta & Denton, 2008). For instance, if an artery blockage is detected, a diagnostic procedure which normally takes 45 minutes may become an interventional procedure that takes twice as long and may cause all subsequent appointments to be delayed. *Second*, emergent patients, with the highest priority, arrive randomly throughout the day and require immediate treatment. This further disrupts the intended flow of operations.

The main performance metrics for a cath lab are idle time of resources, staff overtime, and patient waiting time. Several studies have highlighted the importance of these key metrics for healthcare planning (Cayirli & Veral, 2003; Gupta & Denton, 2008; Gupta et al., 2007; Huang, 1994; Mullen, 2003; Strum et al., 1999). It is important to improve efficiency by minimizing all three metrics. When a cath lab is not utilized during the budgeted time, the lab is being under-utilized and the staff is being paid but no operation is being performed. Also, it is quite possible for labs to be under-utilized and still experience overtime. Ideally, hospital managers would like to avoid such situations. On the other hand, procedures should not be postponed to reduce overtime, since delays in cardiac

catheterizations can lead to patient dissatisfaction and even have negative consequences on patient health (Huang, 1994; Gupta & Denton, 2008; Gupta et al., 2007). Hence, it is critical to improve the efficiency while ensuring the quality of care.

Our research develops a simulation-based framework for analyzing the various factors that affect the efficiency of cath labs in termes of lab utilization, overtime costs and patient waiting times. It is based on real-world data from studying multiple cath labs in a large metropolitan hospital. The factors that we evaluate are size of the time block assigned to each procedure, procedure duration, arrivals of emergent cases, variation in demand as well as the option of rescheduling some patients to the end of the day. The simulation model can be used to develop an efficient frontier, so that a decision maker can easily identify the trade-offs between operating costs, patient waiting time and lab efficiency, and choose the size for the time blocks. The hospital benefited by utilizing the efficiency frontiers generated by the simulation approach in increasing its utilization of cath lab resources by 10%, while reducing overtime by 71%.

The rest of the chapter is structured as follows. Section 2 provides a review of existing literature and section 3 introduces the background of the study. Section 4 describes the simulation model constructed, as well as the design of experiments. Section 5 presents the results from the base model and sensitivity analyses along with the pilot study with our recommended approach and comparisons. Section 6 concludes with directions for future research.

2. Literature review

Discrete Event Simulation (DES) has been extensively used to study health care operations. It allows managers to assess the efficiency of existing health care delivery systems, to ask 'what if' questions and to evaluate managerial alternatives without altering the present system (Jun et al., 1999). An advantage of DES modeling over other mathematical modeling techniques is the ability to precisely capture complex patient flows and then test alternatives by changing flow rules and policies. For example, when emergent patients arrive, a pre-planned sequence of operations may be changed since emergencies must be treated prior to all other patients. DES also has the advantage of easily incorporating variability in interarrival and processing times. Finally, actual data can be easily employed for comparison and sensitivity analysis. In the existing literature (Davies & Davies, 1994; Lowery, 1998), simulation is often the recommended method for modeling health care clinics over analytical and deterministic approaches, mainly, due to the nature and complexity of such systems.

Everett (2002) uses DES to provide decision support for scheduling of elective surgeries in hospitals. Dexter et al. (1999a) use DES to predict the effects of management interventions on decreasing variability in operating room utilization. In a related study (Dexter et al., 1999b), DES is used to model the scheduling of operating rooms to compare and analyze different bin-packing algorithms. In a rolling-horizon environment with varying demand loads, Rohleder and Klassen (2002) use DES to compare different appointment scheduling methods (overtime, double-booking). Romanin-Jacur and Facchin

(1987) uses DES to study the facility dimensioning problem and the sizing of the assistance team in a pediatric semi-intensive care unit. Gupta et al. (2007) study the capacity planning problem in cath labs using DES. De Angelis et al. (2003) interactively use system simulation and optimization to calculate and validate the optimal configuration of servers in a transfusion center. Swisher (2001) develops a DES model to analyze alternatives on staffing levels, facility design scheduling policies and operating hours to see the effects of the changes. In more recent studies, Persson and Persson (2009) use DES to study how health care policies affect the waiting time of patients at a local hospital, and Huang et al. (2009) use DES to evaluate the effectiveness of various planning options and assignment rules for workforce capacity planning. A detailed review of previous research articles on DES in health care is presented by Jun et al. (1999).

The focus of this study was the catheterization facilities within Scottsdale Healthcare (SHC) located in Arizona. SHC had previously implemented lean principles to minimize as much "waste" as possible from their "door-to-balloon" procedures. In spite of the process improvement and standardization, utilization of resources remained low while overtime costs and patient waiting time were rising as the volume of patients was increasing. This led to unsatisfactory operational as well as customer satisfaction metrics. SHC was using block scheduling with block sizes of 120 minutes. Thus, every scheduled procedure was allotted a time block of 120 minutes. When emergent patients arrived, the next free lab was used and the patient previously scheduled was delayed, resulting in a delay for all subsequent cases. An initial investigation showed that this approach was not very

efficient. Most of the procedures were completed well before the allotted 120 minutes and hence the lab and the staff were idle till the start of the next procedure. In most instances, the next procedure could not be advanced since the case start times were assigned earlier and that is when the patient and physician were expected to be ready. Hence, it was essential to develop a framework to analyze the impact of the block size on the cath lab performance. Specifically, the objectives of this study were three-fold:

a. Develop a simulation model that can be used to evaluate the performance of the cath lab.
b. Improve the efficiency of cath labs as measured by (i) lab utilization, (ii) staff overtime and (iii) patient delays.
c. Conduct detailed sensitivity analysis by varying system parameters (such as demand variation, processing time variation) to examine the robustness of recommended block sizes.

In addition, our work adds to existing literature by considering patients with different arrival patterns and priorities in a multi-criteria decision environment, as well as considering the added flexibility of rescheduling patients in order to decrease schedule interruptions and the chain-effects caused by delays or emergencies.

3. Defining patient flow in the cath lab of a hospital

The SHC facility under study has two labs that handle catheterizations. Patients requesting this procedure are classified into three types: (1) *Elective* patients -- These are mostly outpatients that request the procedure at least two

weeks in advance. (2) *Urgent* patients -- These are inpatients that stay in the hospital for other reasons and need a catheterization. Their operation has to be completed within a day of the request. (3) *Emergent* patients -- These are patients that come through the Emergency Department (ED). Their operations have the highest priority and must be performed immediately or as soon as possible. The patient flow in the cath lab is shown in Figure 1.

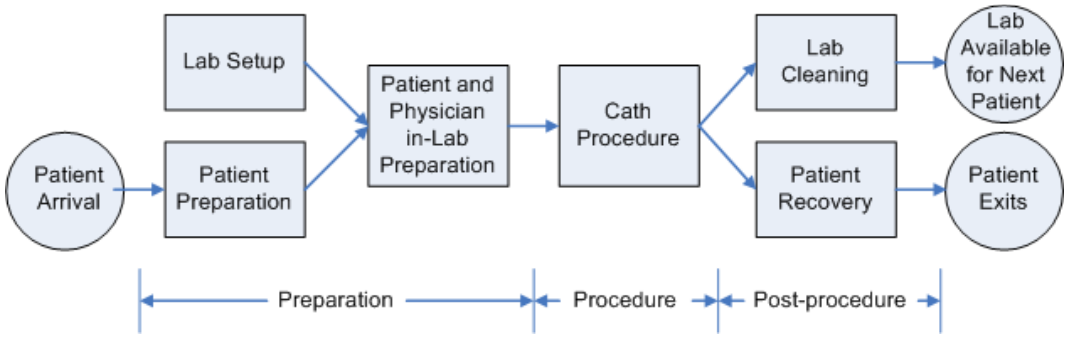


Figure 1. Patient flow through the cath labs

After arrival and admission, the patient is educated about the procedure and the risks associated with cardiac catheterization. At the same time, the staff is setting up the lab by establishing the ECG monitoring and intravenous (IV) access for emergency medications or sedation. The first case of every day requires a little more than 30 minutes of lab-preparation due to equipment and computer start-up and connection to the network. One technician and one nurse arrive 60 minutes prior to the scheduled start time of the first case for the setup. Subsequent clean-up and preparation, referred to as turn-over, require about fifteen minutes. After initial admission procedures, the patient is transferred to the lab where vascular access site preparation and sterile field preparation is performed. After the in-lab preparation, the procedure begins. The duration of the operation can vary from

less than 30 minutes to more than two hours for a variety of reasons, including patient medical history, physician experience and procedure type. After the case is completed, the patient is transferred to a recovery room and the lab is prepared for the next case. The entire process can thus be divided into three phases, namely: preparation, procedure and post-procedure. Preparation and post-procedure are often referred to jointly as *turn-over time*.

As discussed earlier, the SHC cath lab was using block scheduling to develop initial schedules. Each scheduled procedure was assigned 120 minutes. Elective and urgent cases are scheduled ahead of time. If an emergent case arrives, either the lab that is free or the next available lab is used. The procedure previously scheduled in that lab is postponed and the patient is delayed. This also results in delaying all subsequent cases in that lab. This will be discussed further in section 4.1. Each of the labs worked for nine hours with a 30-minute lunch break. The starting times of the labs were staggered by 30 minutes to avoid congestions. Cases that required time beyond the nine-hour regular shift were completed using overtime labor.

When we analyzed the history of past cases, three things emerged: *First*, utilization of the labs (i.e., the percent of time that they were being used during regular hours) was only 43% on average. *Second*, staff often had to work overtime (about 353 minutes on average per week) to complete the cases scheduled during the day. *Third*, patients were experiencing long waiting times. It is interesting to point out that the hospital had a low utilization of the cath labs and high level of overtime. This clearly indicated inefficiencies in patient scheduling since SHC

had implemented several lean principles to standardize many of the cath lab's controllable operations. Hence, management wanted to investigate how to better schedule patients in order to balance utilization of resources, overtime, and patient waiting times and improve customer satisfaction.

4. Methods and analyses

As a first step toward understanding the process, we collected data on all the procedures completed in the two labs for a period of 6 months (October 1st, 2006 to March 31st, 2007). This included the busiest season of the year. A preliminary analysis of the data showed that on average, there were four scheduled elective cases and three scheduled urgent cases per day. In the peak season, which is December, January and February, there were six elective cases and four scheduled urgent cases per day. Random arrivals of emergent patients had a Poisson distribution with a mean of 2.5 patients per week. Using the historical data, we statistically fit probability distributions to describe the three phases of the operation. We also collected data on physician lateness and incorporated it as part of the preparation time. Finally, we aggregated the three phases to determine the total case duration and fit a distribution for this as well. Table 1 presents a summary of the distributions for these phases and the total case duration.

Table 1

Probability distribution of case duration

	Distribution	*Mean (min)*	*Std. Deviation (min)*
Preparation (incl. Physician Lateness)	Erlang	23	13
Procedure	Beta	42	32
Post-procedure	Lognormal	8	5
Total case duration	Gamma	73	36

Both elective and urgent cases are scheduled a day before the surgery. We use block scheduling to generate an initial arrangement for elective and urgent cases. To provide some safety cushion for the variation in case duration as well as the arrival of emergent cases we adjust the schedule in three ways. *First*, for each case scheduled we allocate a time window that is larger than the historical median. *Second*, a buffer is added to the lunch break to decrease the effect of morning delays on the cases that follow in the afternoon. *Finally*, idle time is allocated at the end of the daily schedule to decrease the possibility of overtime. A sample initial schedule with 10 patients per day, 90 minutes allocated per case and 30-minute buffer is shown in Figure 2.

We use Arena 10.0 to model the patient flow through the two cath labs. Generalized capacity planning models often assume that the current resources are achieving maximum capacity (VanBerkel and Blake, 2007). We assume that there are 10 patients scheduled per day and perform sensitivity analysis on the demand. Patients are assumed to be punctual. We treat elective and urgent patients the same in this study, because data analysis did not provide statistical evidence that

there is a difference in case duration. All cases are allocated the same length of time regardless of the type of procedure for the ease of implementation.

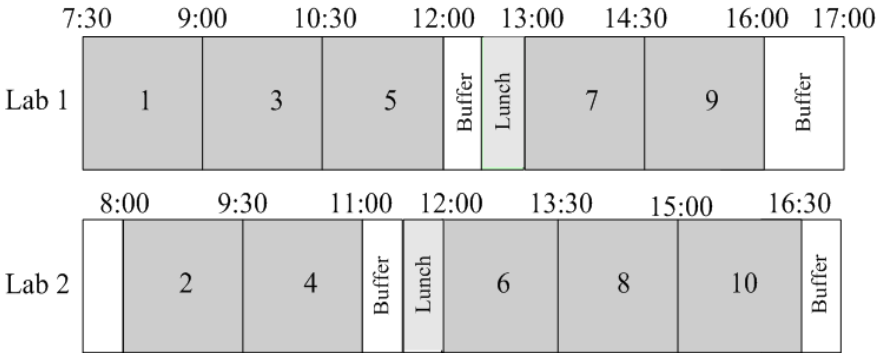


Figure 2. Schedule of cases with allocation of 90 minutes

An entity in the model corresponds to a patient arriving and capturing available resources, i.e. cath labs. Patient-arrival is a model input and arrivals occur exactly as scheduled. Upon arrival, if a lab is available, the case starts immediately. Otherwise, the scheduled and emergent cases will wait for the first available lab. The emergent cases have the highest priority and the other cases are scheduled in the order of arrival. Once a case is assigned to a lab, it occupies it for the duration of time sampled from the three distributions, respectively, as shown in Table 1. Upon case completion, the lab becomes available for the next patient.

The model output captures the resulting lab utilization, overtime incurred and patient waiting times in each scenario. Utilization is defined as the fraction of the budgeted time that the lab is being utilized. When the lab is not utilized during regular hours, the lab crew still gets paid. Hence, under-utilization can also be translated into a direct cost measure as [(1–utilization) × regular salary for the lab crew]. Overtime is defined as the time the staff is working after the budgeted time.

This has a direct cost implication and can be captured using total overtime cost, calculated as (overtime salary for the lab crew × total overtime). Exact regular and overtime salary rates were provided by SHC. Patient waiting time is defined as (max{0, (actual start time – scheduled start time)}). Since two of the performance measures have been translated to costs, we transformed the problem into developing a schedule that minimizes two criteria: total cost of overtime and under-utilization and total patient waiting time.

4.1.Rescheduling

In order to reduce the adverse effects of delays on scheduled cases caused by emergent arrivals and process variation, we consider the option of rescheduling inpatients to the end of the day. Since these patients are already in the hospital, they can be taken back to their room and brought to the cath lab later in the day for the procedure. A patient may be rescheduled for two reasons, (i) due to an emergent case arrival, and (ii) when a patient has been waiting longer than a predetermined time. However, to maintain service quality and patient satisfaction, we use the following constraints. First, to ensure patient safety, an emergent case will not be rescheduled. Second, to ensure patient satisfaction, a case cannot be rescheduled more than once. Finally, elective cases will not be rescheduled, since these are outpatients.

The hospital decided not to reschedule patients. Based on our observations and interviews, this is due to the fact that the feasibility of rescheduling depends on physician availability. Most patients are assigned to a specific physician and re-assigning on short-notice is challenging. However, the administration also

indicated that it might be better to work with the physician and reschedule the patient rather than have the physician wait in the hospital for a cath lab. Considering this, we have experimented both with and without the option of rescheduling, and compared the results.

4.2.Parametric analysis

Our goal in using the simulation model is to understand the impact of critical decision variables on the output metrics. Hence, we conducted detailed experiments using the following decision variables: (a) time-block length L, (b) patient waiting time before rescheduling W, and (c) lunch buffer length B. For our parametric analysis the duration of the time-block (L) was based on percentiles of the case duration distribution, which was derived from historical data. Specifically, the scheduling approach used at SHC allocated two hours per case (i.e. L = 120 min) which corresponds to the 92nd percentile of the total case duration distribution. The time-block lengths used for our parametric analysis ranged from the 55th to the 95th percentile and are presented in Table 2. The waiting time of patients before rescheduling (W) was evaluated at 30%, 40%, 50%, 60% and 70% of the time-block length (L). We tried 45-minute and 30-minute lunch buffers (B) for lab 1. Lab 2 can only have a 30-minute lunch buffer since the lab starts 30 minutes later than lab 1. Using a factorial design of experiments, we generated a total of 120 experimental scenarios (10 time-block lengths × 5 rescheduling wait times × 2 lunch buffer combinations) for schedules with rescheduling and 20 scenarios (10 time-block lengths × 2 lunch buffer combinations) for schedules without rescheduling. Each experimental scenario was simulated for 100 days.

Table 2

Time-block lengths used in the parametric analysis

Percentile of case duration	55	60	65	70	75	80	85	90	92	95
Time-block length (min)	65	70	75	80	85	90	100	110	120	140

4.3.Experimental results

Figure 3 presents the results of the simulation experiments for schedules that allow waiting patients to be rescheduled. The results indicate that scenarios with time-block length below the 60^{th} percentile and above the 85^{th} percentile are clearly dominated. In order to enhance clarity, we are not considering those in the following analyses. For each combination of L and B, there are five points on the graph corresponding to the 5 different values of W. The horizontal axis contains the average weekly under-utilization and overtime cost and the vertical axis contains the average weekly total waiting time. Each point in the graph represents a combination of the three decision variables (L, W, B).

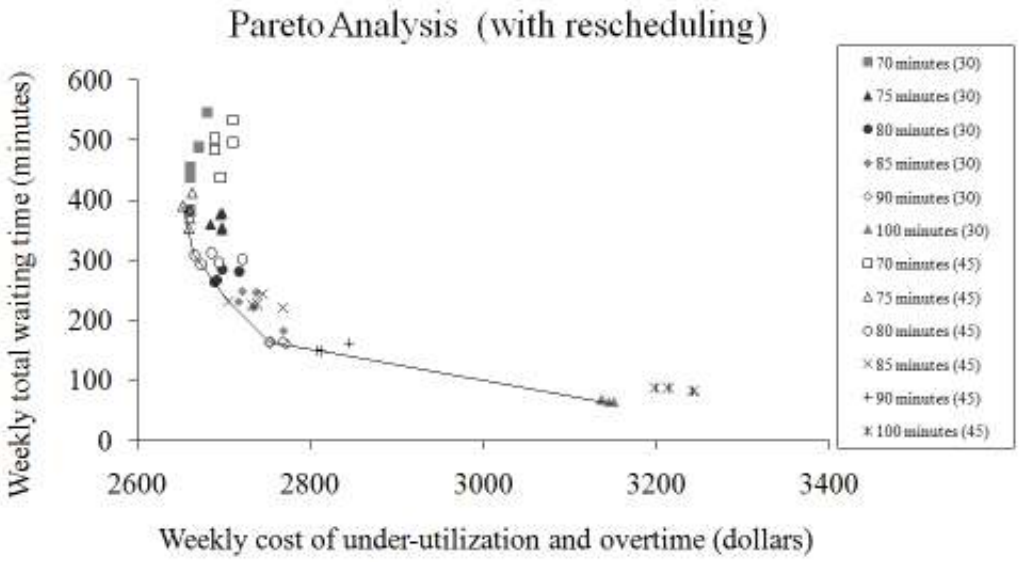


Figure 3. Results with rescheduling

These results demonstrate that larger (smaller) time blocks result in shorter (longer) patient waiting time and higher (lower) total costs. For instance, Allocating 70 minutes per case yields a weekly cost of $2710 and waiting time of 533 minutes. However, allocating 100 minutes per case dramatically reduces the waiting time but increases the cost. The graph provides the efficient frontier (shown with the solid curve) for the managers to trade-off between the operational cost and patient waiting time. Based on the feedback from staff and management, we suggested a 90-minute time block allocated per case with a 45-minute lunch buffer for lab 1 and 30-minute lunch buffer for lab 2**,** and 55-minute waiting before suggesting rescheduling an inpatient, as marked on Figure 3. The weekly cost and patient waiting time at this level are $2753 and 163 minutes, respectively.

Figure 4 presents the results of the simulation analysis without rescheduling patients. Once again, scenarios that were clearly dominated have

been omitted from the graph. Based on the results, we suggested a time-block length of 90 minutes per case, a 45-minute lunch buffer for lab 1 and 30-minute lunch buffer for lab 2. The weekly cost and patient waiting time at this level are $3807.34 and 108 minutes, respectively.

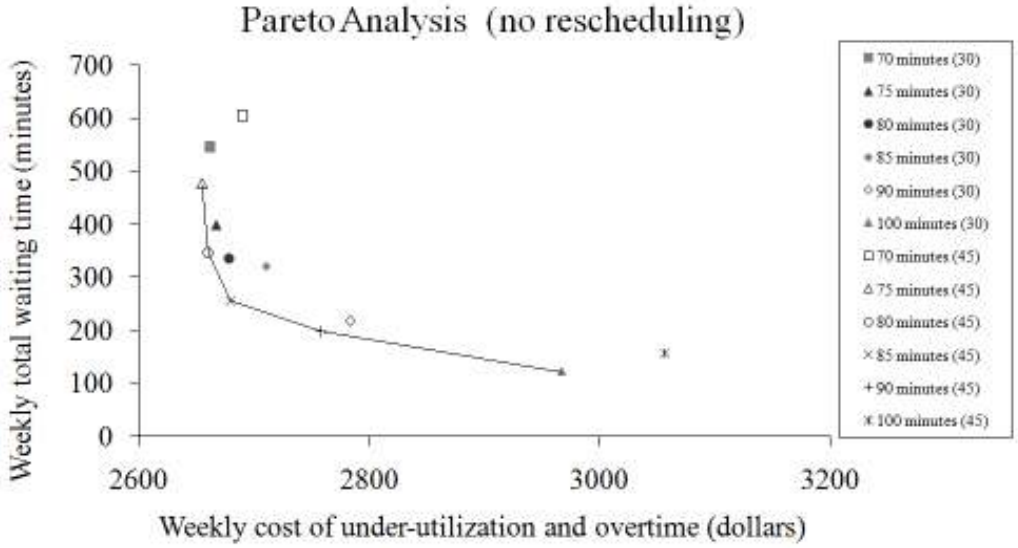


Figure 4. Results without rescheduling

5. Comparison with historical data

In order to validate the recommended time-block length, we collected data on the actual number of patients per day, case duration, and time of emergent patient arrivals for the first eight weeks of 2007 (January 1st to February 23rd). We scheduled the same set of patients using the recommended time-block length and lunch buffers, considering both with and without rescheduling of waiting inpatients. The results are presented in Tables 3 and 4. Table 3 indicates that from an operational perspective, our recommendations would have increased the average utilization by approximately 26%. The total overtime in eight weeks was

significantly reduced from 2738 minutes to 297 (with rescheduling) and 201 (without), a reduction of approximately 90%.

Table 3

Comparison with SHC approach – Operational factors

		Average utili-zation	*Total over-time (min)*	*Number resche-duled (min)*	*Total cost of overtime and under-utilization ($)*
SHC Approach		43.60%	2738	N/A	13,643.11
Simulation Results of SHC Approach		51.22%	1168	0	12,342.09
Recommended Approach	With rescheduling	69.31%	297	3	4,704.73
	Without rescheduling	69.72%	201	0	4,432.54

Prior to our study, data on the waiting time of individual patients was not collected. However, according to the perception of nurses, patients were experiencing excessive waiting times. Table 4 compares the total waiting time incurred by patients when using the recommended schedule with and without rescheduling. As expected, rescheduling improves both the average waiting time and the possibility of a patient having to wait. All measures, (i.e. percent of patients waiting, average waiting time and total waiting time) were within the hospital's acceptable range.

Table 4

Comparison with SHC approach – Patient waiting times

		Percentage of waiting patients	*Average waiting time (min)*	*Total waiting time (min)*
SHC Approach		N/A	N/A	N/A
Simulation Results of SHC Approach		16.26%	15.79	521
Recommended Approach	With rescheduling	22.36%	19.13	880
	Without rescheduling	25.13%	21.16	1080

Overall, these results indicate that a sensible adjustment of the time allocated to each case and the addition of small buffers to allow for uncertainty and variation can improve the performance of labs, reduce cost and significantly decrease overtime.

5.1.Sensitivity analysis on scheduling performance

In order to test the robustness of the model and the extent to which these results can be generalized, we performed sensitivity analysis on parameters that affect scheduling performance. Specifically, we considered the case duration, the demand for elective and urgent cases, and the emergent patient arrivals, as these are key factors that influence the schedule and overall efficiency and utilization of the cath labs.

Case Duration Distribution: We first wanted to understand the impact of the case duration distribution on the schedule. Previous studies also show that a lognormal distribution is usually a very good fit for capturing the variations and uncertainties inherent in surgical procedure durations (Kaandorp & Koole, 2007).

Our initial study used a Gamma distribution. In order to explore this further, we collected additional case duration data from two different SHC facilities. We analyzed the 2048 cases and determined that the Lognormal distribution described the data very well. This, along with results from published studies that supported our conclusion, motivated us to assume that the case duration can be described by a Lognormal distribution in all of our sensitivity analyses.

Table 5 presents the percentiles from the 55th to the 90th for the Lognormal case duration distribution obtained with the 2048 cases. For each time-block, we ran the simulation for 100 days. Pareto analysis showed that the 70th and 75th percentiles (90 and 95 minutes, respectively) are the most desirable options in terms of minimizing both waiting time and cost of overtime and under-utilization. Processing time varies by case, patient and physician. We performed sensitivity analysis by varying the coefficient of variation of the process duration from 0.1, 0.25, 0.5, 0.75 and 1. Pareto analysis showed that the 75th percentile remains on the efficient frontier for all ranges.

Table 5

General data percentiles

Percentile	55	60	65	70	75	80	85	90
Procedure time (min)	75	80	85	90	95	105	115	130

We conjecture that the 75th percentile will, in general, be on the efficient frontier for most scenarios. In the following sections, we study the impact of variation in demand for elective and urgent cases, as well as variation in emergent case arrival on the Pareto-optimal time-block length.

Demand for Elective and Urgent Cases: Literature shows that heart disease cases show a winter peak (Spencer et al., 1998), especially in Arizona because retired people move here in the winter. In order to test if the 75th percentile would be Pareto-optimal during peak seasonal demand, as well as lower demand, we varied the demand to capacity ratio from 0.7 to 1.3, in intervals of 0.1. A ratio of 1.3 indicates that demand is 30% more than available capacity and a ratio of 0.7 indicates that demand is 70% of capacity. For each ratio, we used eight different time-block lengths as shown in Table 6. Every ratio - time-block length combination was simulated for 100 days. Pareto analysis showed that the 75th percentile was on the efficient frontier for all levels.

Specifically, we saw that longer time blocks (80th-85th percentile) tend to perform better when the demand-to-capacity ratio is low. In these cases, the total waiting time is lower without a significant increase in the overall cost. For example, when demand is 70% of capacity, using the 75th percentile yields weekly cost of $3,746 and a total waiting time of 146 minutes, while the 85th percentile generates weekly cost of $3,849 but with a total waiting time of 63 minutes. From a management point of view however, we do not recommend increasing the block size during off-peak seasons as this may give physicians and staff the impression that there is more than enough time and lead to inefficiencies, as the demand-to-capacity ratio starts increasing. Similarly, when the demand-to-capacity ratio is high, shorter time blocks may be preferred as they reduce the total cost without significantly increasing the waiting time. For example, by

moving from the 75th to the 70th percentile, the weekly cost is reduced by $262 while the total weekly waiting time increases by 86 minutes.

Cath lab managers ideally would like to keep the time-block length constant throughout the year. This would make planning and scheduling consistent and less cumbersome. In view of this and the results obtained from our experimentation, the 75th percentile of the case duration distribution seems to be the logical choice for the time-block length.

Arrival of Emergent Cases: Emergencies are a random and critical part of demand that affect the schedule dynamically. Not surprisingly, data analysis shows that emergent arrivals follow a Poisson distribution. We perform sensitivity analysis by changing the coefficient of variation ($CV = 1/\sqrt{\lambda}$) from 0.25, 0.5, 0.65, 0.75, 0.9 and 1. Each scenario was run for 100 days. Pareto analysis using simulation for the eight percentiles (from 55th to 90th) showed that the 75th percentile remains on the efficient frontier at all levels of CV.

In conclusion, sensitivity analysis on demand, emergent arrival variance and procedure duration variance, shows that using the 75th percentile of total case duration as a general rule, is overall an efficient and reasonable choice as it balances all aspects of performance in healthcare scheduling.

5.2.Implementation and comparison

SHC has implemented our recommendations since January, 2008. We use the same eight-week data in 2008 as in 2007 to compare the performance. Results are presented in Table 6.

Table 6

Comparison of performance before and after implementation

Metrics	*2008*	*2007*
Utilization	52.03%	43.60%
Avg. weekly overtime (min)	98	342
Total patient waiting time (min)	1609	N/A
Total Number waited	72	N/A
Avg. waiting time per patient (min)	22.35	N/A
% of patients waiting	33.33%	N/A
No. of cases	216	207
Total case duration (min)	17991	16973
Total cost of over-/ under-utilization ($)	8,090.10	13642.42

The improvement in utilization is less than what was predicted by simulation, because in 2008 the number of cases and the total case duration is different from 2007. However, considering both under-utilization and overtime costs, the savings for these eight weeks were $5,552. Coincidentally, we also find that physician lateness, which significantly contributes to preparation time variation, is reduced in 2008. This was reflected by two facts: (1) the number of cases with physicians' lateness is reduced; (2) the length of lateness is proved to have been statistically reduced by a t-test with 99% confidence. This may be due to the sense of urgency created by the shorter time-block allotted to each case.

Table 7

Physician lateness data in 2007 and 2008

Physician Lateness	Mean	Variance	Frequency
2008	10	107	159
2007	14	128	186

6. Conclusions

In this chapter we have developed a simulation model to evaluate the efficiency of cath lab operations while varying key parameters such as length of the time-block assigned to each case, length of lunch buffers as well as the option of rescheduling patients. Our analysis considers both operational costs and patient satisfaction metrics and illustrates the tradeoffs between the two. Detailed experimentation has helped recommend allocating to each case a time block equal to the 75^{th} percentile of the case duration distribution and schedule a short buffer in the middle and at the end of each day to absorb variation and reduce the possibility of overtime.

In order to test the robustness of our recommendations we perform sensitivity analysis on key variables including demand, process duration, emergent case arrivals and also combine data for the busiest months from two separate locations and compared the results. Overall we find that the 75^{th} percentile of process duration is always on the efficient frontier and is a good compromise of both operational cost and patient waiting well. The health care facility adopted our recommendations and is now realizing the anticipated improvements. An interesting extension of this study would be considering physician specific data such as differences in lateness and/or average case duration. Incorporating this information in the analysis while developing the initial schedule may further improve performance, both in terms of efficiency and patient satisfaction.

CHAPTER 3

A TWO-PHASE MULTI-CRITERIA APPROACH FOR THE OPERATING ROOM PLANNING AND SCHEDULING PROBLEM

1. Introduction

Between 1999 and 2007 in the United States, healthcare consumed 35.7% of the real increase in per capita income, and the share of US Gross Domestic Product (GDP) devoted to healthcare rose from 13.7% to 16.2% (Chernew et al., 2009). A report forecasts that the healthcare costs could rise to 34% of GDP in three decades unless something is done to overhaul the industry ("Moving up", 2009). Hospitals face an increasing pressure for efficient resource usage and high quality care in such environment. Surgery accounts for 40% of a hospital's resource costs (Macario et al., 1995), with personnel, infrastructure, equipment, logistics and administrative support costs accounting for most of this cost and material cost being smaller (Roland et al., 2006). Since the OR is one of the key hospital resources, there should be efforts to continuously lower cost and enhance quality. At the same time, planning and scheduling the OR is challenging due to conflicting priorities (Glouberman & Mintzberg, 2001; Ozcan, 2005), internal and external uncertainties (Gupta, 2007), and scarcity of costly resources.

The objective of this study is to address the following problems through developing a concrete model for the strategic level of OR planning and scheduling.

a. How should hospitals allocate OR time to surgical specialties and groups?

b. How should hospitals assign and schedule patients considering both cost efficiency and patient satisfaction?

The mathematical model and algorithms we describe in this study aim to investigate answers to the following important questions:

a. Does the size of a surgical group affect scheduling performance? If so, how much is the impact?
b. How much interaction is there between the two phases of this problem? In other words, what is the impact of decomposing the problem to a planning phase and a scheduling phase?

The rest of the chapter is organized as follows. The next two sections present an introduction to OR scheduling and an overview of previous literature in this area, respectively. In section 4, the problem is then modeled in two phases, advance scheduling phase and allocation scheduling as mixed integer programs (MIP). We consider multiple objectives in each phase and we investigate the impacts of the decomposition of the problem. In this section, we also develop heuristics and a Random Keys Genetic Algorithm for daily patient scheduling problem of the second phase. Experimental results are then presented in section 5. Finally, we discuss our conclusions and future research directions.

2. Problem description

Three classes of patients are generally considered in OR planning and scheduling: elective, urgent and emergency patients. Elective surgeries are usually requested a few weeks in advance. On the other extreme, emergency patients need to be immediately performed. Urgent patients are sufficiently stable so that they

can be postponed for a short period, i.e. a few days. Urgent and emergency patients sometimes are classified together as non-elective patients. OR departments can plan for non-elective surgeries ahead of time by reserving partial OR capacity. The reserved capacity may be concentrated in an OR (ORs) that is (are) entirely dedicated for non-elective surgeries. However, this usually leads to low utilization in the dedicated rooms. Another way is to allocate the slack to a number of ORs scheduled with elective surgeries, allowing non-elective surgeries to be scheduled in between two elective surgeries. In this study, we adopt the second option, which is to reserve some capacity in every room for non-elective surgeries. The actual arrival and duration of emergency patients will not be considered in this study.

Surgical cases have three stages: preoperative, perioperative and postoperative (Pham & Klinkert, 2008). In the preoperative stage, patients get necessary preparations, including certain instructions, paperwork, medication, etc., and then are moved to an OR. In the second stage, patients are anaesthetized and surgeries are performed. In the last stage, patients are transported to PACU to recover. In this study we do not consider the first stage, because 1) preparation procedures are usually quite standardized and do not have much uncertainty; 2) the arrival time of elective and urgent patients, which make up 90% of all patients, are scheduled. The capacity planning and staffing of the preoperative stage can thus be well determined ahead of time. However, the duration of the second stage is not as predictable and all patients from different ORs all share PACU resources in the third stage. If there is no available bed in PACU when the surgery

completes in OR, the patient may be held in the OR until a PACU bed is available. It is considerably more costly for patients to recover in the OR.

The most popular basic OR scheduling approaches are open scheduling and block scheduling (Ozcan, 2005). Open scheduling allocates surgery times to the first surgeon requesting them. A limit on the number of times allocated to that surgeon, or to the estimated surgical time may be imposed. This approach has several critical drawbacks, such as simultaneous OR overtime and idle time, and high cancellation rates due to overbooking (Ozcan, 2005). With block scheduling, a block of OR time, usually one-half to a full day, is allocated exclusively to a surgical group, which is composed of one or multiple surgeons in the same specialty. Based on the availability of surgeons and historical demand patterns, a "master schedule" is first developed with surgical groups assigned to one to two week repeating time blocks until there are major changes in demand or surgical groups (Roland et al., 2006). Table 8 shows an example of OR block allocation for a 2-OR hospital with 20 blocks allocated to surgical groups, assuming there are two ORs with 20 blocks, and four surgical specialties in the hospital. The advantage of the block system is that it increases utilization through better afternoon usage of the OR. It also guarantees surgeons surgical times and allows them to know surgical start times well in advance (Ozcan, 2005).

Table 8

An example of surgical block scheduling

	OR 1		*OR 2*	
	8:00-12:00	*13:00-17:00*	*8:00-12:00*	*13:00-17:00*
Mon	Anesthesiology	Oral maxillofacial	Urology	Urology
Tue	Anesthesiology	Anesthesiology	Oral maxillofacial	Urology
Wed	Oral maxillofacial	Ophthalmology	Oral maxillofacial	Oral maxillofacial
Thu	Urology	Urology	Anesthesiology	Anesthesiology
Fri	Ophthalmology	Ophthalmology	Anesthesiology	Ophthalmology

In practice, block scheduling is divided into three sub-procedures. Firstly, a master cyclic operation schedule is developed and surgical groups are assigned to blocks. Secondly, elective patients are assigned time blocks and surgical groups according to the availability of recourses. After blocks are assigned, the sequence of patients is determined. The first and second step together are referred to as "advance scheduling" (Magerlein & Martin, 1978) in the literature. Thirdly, patients are sequenced on each day of surgery, with considerations of urgent patients. This step is referred to as "allocation scheduling" (Magerlein & Martin, 1978). In this study, we develop our approach based on such a two-phase structure.

3. Literature review

In the advance scheduling phase, budgets often determine the total OR time available, and there are several factors that determine the proportion of time to be assigned to each surgical specialty, such as waiting times (Dexter & Traub,

2002), OR efficiency (Dexter & Traub, 2002), and equity among all the specialties (Blake & Dexter, 2002). Ogulata and Erol (2003) develop a set of hierarchical multiple criteria mathematical programming models to generate weekly operating room schedules. The objectives considered are maximizing utilization of operating room capacity, balancing distribution of operations among surgeon groups and minimizing patient waiting times. Marcon et al. (2006) model the problem as a multiple knapsack problem while minimizing the difference of workload between rooms and minimizing the risk of no-shows. The allocation scheduling phase is more operationally focused. Ozkarahan (2000) uses goal programming to assign cases to ORs in order to minimize the sum of ORs' undertime and overtime costs, and then sequences the loaded cases according to some priority rules. Pham and Klinkert (2008) use an MIP model formulation to minimize the weighted sum of makespan and the starting times of all surgeries. They also propose that add-on and emergency surgeries can be scheduled by adding new constraints using job insertion. Jebali et al. (2006) develop a two-step MIP formulation considering both phases. Fei et al. (2006) also develop a MIP model and solve the two-phase scheduling problem by column generation. Cardoen and Demeulemeester (2007) use simulation to tackle the problem and they include overtime and patient waiting time in their evaluation criteria.

In this study we consider both phases with multiple objectives in each phase. Although this problem has attracted much attention, there are still some open challenges that need more attention. Firstly, much previous research is concerned with patients' waiting time on the day of surgery (Cardoen &

Demeulemeester, 2007; Jebali et al., 2003), while the time they spend in the wait list to be scheduled is more important from a patient safety perspective (Marcon & Dexter, 2006). Secondly, many studies focused solely on the OR (Jebali et al., 2003; Gerchak et al., 1996; Hans et al., 2008), however, other parts of the surgical suite can have an impact on the performance as we discussed in section 2. Lastly, the vast majority of the literature tries to optimize OR scheduling by splitting it into a planning phase and a scheduling phase (Fei et al., 2006; Hans et al., 2008), and each phase is considered separately. However, the two steps interact with each other in reality and a bad assignment in the planning phase may influence the performance of the scheduling phase (Roland et al., 2006).

Table 9 shows papers that consider both advance and allocation scheduling. In the header row, the most commonly used objectives in the literature and in hospitals are listed. As is shown in the table, recent papers tend to consider multiple objectives.

The models proposed in this paper take into account all the objectives except the patients' waiting time on the day of surgery because we assume surgery durations are known ahead of the time as we illustrate in section 4. We will also study the importance of considering both phases interactively. To the best of our knowledge no surgery scheduling models have been proposed that consider all these perspectives.

Table 9

Summary of literature considering both phases

Author (Year)	Sur-geon prefer-ences	Patients' waiting on the waiting list	Patients' waiting on the day of surgery	OR Utiliza-tion	Over-time	Level-ing PACU	Com-ments
Lowery *et al.* (1999)				S			
Vandan *et al.* (2000)			S		S		
Jebali *et al.* (2003)			M	M	M		
Everrett (2004)		S		S*			*Ward utilize-tion
Sciomachen *et al.* (2005)		S			S		
Sandberg *et al.* (2005)				A			
Fei *et al.* (2006)				D,C,T	D,C,T		
Krempels *et al.* (2006)	H		H				
McIntosh *et al.* (2006)				H,A	H,A		
Roland *et al.* (2006)			M,G		M,G		Inter-action
Jebali *et al.* (2006)		M		M	M		
Cardoen *et al.* (2007)		S	S		S	S	
Gupta (2007)	SP		SP	SP	SP		
Testi *et al.* (2007)	M,S	M,S	M,S	M,S	M,S		
Gupta *et al.* (2007)	S		S	S	S		
Hans *et al.* (2008)		S,SA		S,SA	S,SA		
This study	X	X		X	X	X	Inter-action

(*S: simulation, M: mixed integer programming, A: qualitative analysis, D: dynamic programming, C: column generation, T: tabu search, H: heuristics, G: genetic algorithm, SP: stochastic programming, SA: simulated annealing*.)

4. Solution approaches

4.1.Model description and assumptions

In this research we use block scheduling and develop a two-phase scheduling approach. In the advance scheduling phase (Phase 1), the surgical groups are first assigned time blocks (Phase 1.1). This is done yearly or seasonally depending on the variation in demand, and the block schedule will repeat every one or two weeks. The allocations of blocks to surgeons are revisited when there is a change in capacity, number of surgeons, or when medical technology innovations alter the capacity usage of certain types of procedures (Gupta, 2007). The objective is to satisfy surgeons' preferences as much as possible. Next, patients are assigned to surgical groups and time blocks (Phase 1.2); this takes place every one or two weeks at the beginning of each cyclic period. The objectives are to minimize patients' waiting on the waiting list, under-utilization and overtime in OR. In the allocation scheduling phase (Phase 2), the goal is to find the optimal sequence for all patients each day and the objectives are to minimize the overtime and to regulate the patient flow from OR to PACU by minimizing the maximum number of beds in PACU in use at any time. The lack of PACU beds may lead to OR blocking (as discussed in section 2) and the staffing cost in PACU is determined by peak demand. Figure 5 shows the planning horizon and objectives of each phase.

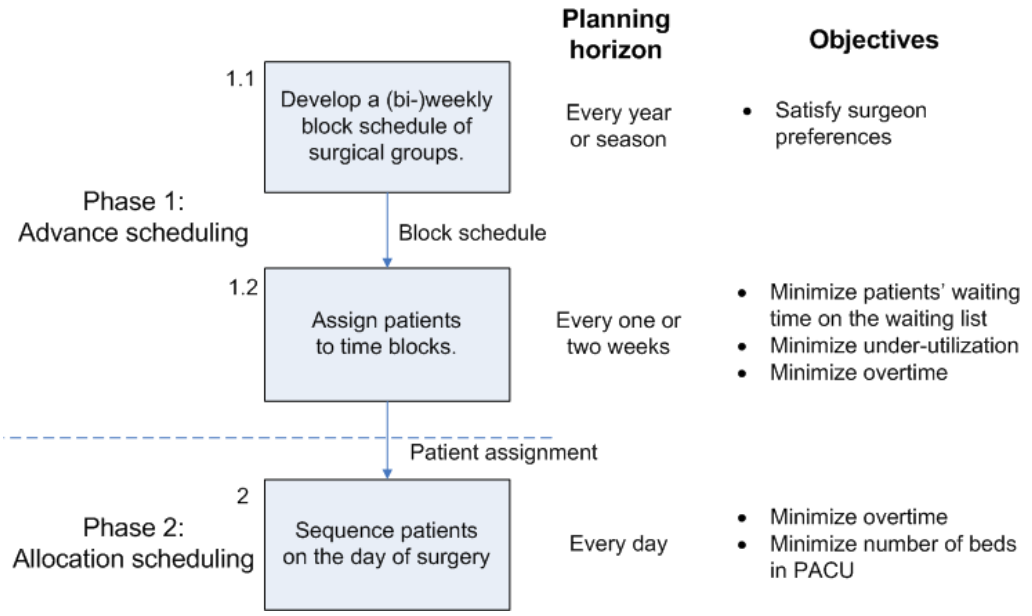


Figure 5. Illustration of the two phases in our approach

We model each step with a mixed integer program. The assumptions are:

1) Each day is divided into two blocks (a morning and an afternoon block) with a lunch break in between.
2) Cases that cannot be finished in the morning will use the time in the lunch break.
3) Cases that cannot be finished during regular hours and the lunch break will be pushed to overtime. By law, overtime cannot exceed 4 hours.
4) If there is no available bed in the PACU when a case finishes in OR, the patient has to recover in OR until there is an available bed in PACU or fully recovered.
5) A user-specified capacity in each room is reserved for emergency surgeries.
6) Case durations and recovery times are known.

7) Surgical demand is not greater than the capacity.

4.2.Mathematical model

Decision variables are in uppercase letters throughout this dissertation. Furthermore, the results of some decision variables would feed the next phase, such as the decision variable of block allocation to surgical groups in Phase 1.1, and they will be changed to lowercase in the next phase but the same notation is kept for consistency. In Phase 1, the planning problem is to allocate surgeons and patients with time blocks.

Notation in Phase 1:

$n \in N$	index of room
$s \in S$	index of specialty
$m \in M$	index of surgical group
$r \in R$	index of surgeon
$p \in P$	index of patient
$t \in T$	index of time block (2 blocks per day)
$surg_{rm}$	1 if surgeon r is in surgical group m, 0 otherwise
$surs_{ms}$	1 if surgical group m is in specialty s, 0 otherwise
ra_{nt}	1 if room n is available in block t, 0 otherwise
rs_{ns}	1 if room n can be assigned to specialty s, 0 otherwise
τ_p	estimated surgery time of patient p
dd_p	due date of patient p
pts_{ps}	1 if patient p is in specialty s, 0 otherwise

pg_{pr} 1 if patient p can be assigned to surgeon r, 0 otherwise

sa_{rt} 1 if surgeon r is available in block t, 0 otherwise

pr_{mt} 1 if surgical group m does not prefer block t, 0 otherwise

brk_t length of the lunch break after block t

cap_t capacity of block t per room

e_t emergency demand in block t (*from historical data*)

ut target block usage in an open OR

ot maximum possible overtime per room per day

α_1 relative weight factor of objectives in model 1.2

c_1, c_2 cost of under-utilization and overtime, respectively

The objective of Phase 1.1 is to find a cyclic operation schedule. The decision variables include:

X_{nst} 1 if specialty s is assigned room n in block t, 0 otherwise

Y_{nmt} 1 if surgical group m is assigned room n in block t, 0 otherwise

MIP formulation of phase 1.1:

$$\textbf{Min} \sum_m \sum_t \left(pr_{mt} \sum_n Y_{nmt} \right) \quad (1)$$

$$\sum_s X_{nst} \le ra_{nt} \qquad \forall n \in N, t \in T \quad (2)$$

$$\sum_t X_{nst} \le rs_{ns} \qquad \forall n \in N, s \in S \quad (3)$$

$$X_{nst} \le \left\lceil \frac{\sum_m \left(\sum_r (sa_{rt} \cdot surg_{rm}) \cdot surs_{ms} \right)}{\sum_m \left(\sum_r (sa_{rt} \cdot surg_{rm}) \cdot surs_{ms} \right) + 1} \right\rceil \qquad \forall n \in N, s \in S, t \in T \tag{4}$$

$$Y_{nmt} \le \sum_s (X_{nst} \cdot surs_{ms}) \qquad \forall n \in N, m \in M, t \in T \tag{5}$$

$$\sum_n Y_{nmt} \le \sum_r (surg_{rm} \cdot sa_{rt}) \qquad \forall m \in M, t \in T \tag{6}$$

$$X_{nst}, Y_{nmt} \in \{0,1\} \qquad \forall n \in N, m \in M, s \in S, t \in T \tag{7}$$

The objective (1) minimizes the dissatisfaction of surgical groups. Constraints (2) ensure that each OR is assigned to at most one specialty at one time only if the room is available. Constraints (3) indicate that each OR may be assigned to one specialty at one time only if the room can be assigned to that specialty. Constraints (4) ensure that each OR is assigned to at most one specialty at one time only if at least one surgeon of the specialty is available in that time block. Constraints (5) guarantee that a surgical group is assigned to an OR at one time only if their specialty is assigned with the time block. Constraints (6) indicates that the amount of ORs assigned to each surgical group in a time block has to be at most the number of surgeons available in that block. Constraints (7) ensure all the variables in this model are binary.

Phase 1.2 is to assign patients to time blocks. The decision variables are:

Z_{rpt} 1 if patient *p* is assigned to surgeon *r* on block *t*, 0 otherwise

U_{nt}^{-} under-utilization in room *n* on block *t*

U_{nt}^{+} over-utilization in room *n* on block *t*

W_p waiting time of patient *p* on the waiting list after the due date.

The result of Y_{nmt} from Phase 1.1 is input to Phase 1.2 and presented as y_{nmt}.

MIP formulation of phase 1.2:

$$\textbf{Min } \alpha_1 \cdot \sum_n \sum_t \left(c_1 \cdot U_{nt}^{+} + c_2 \cdot U_{nt}^{-}\right) + (1-\alpha_1) \cdot \sum_p W_p \tag{1}$$

$$Z_{rpt} \le \sum_s \left(pts_{ps} \cdot \sum_n \sum_m \left(y_{nmt} \cdot surs_{ms} \cdot surg_{rm} \cdot sa_{rt}\right)\right) \qquad \forall r \in R, p \in P, t \in T \tag{2}$$

$$\sum_t Z_{rpt} \le pg_{pr} \qquad \forall r \in R, p \in P \tag{3}$$

$$\sum_r \sum_t Z_{rpt} \le 1 \qquad \forall p \in P \tag{4}$$

$$\sum_m \left(y_{nmt} \sum_p \sum_r \left(Z_{rpt} \cdot \tau_p \cdot surg_{rm}\right)\right) + e_t \le cap_t + brk_t + \frac{ot}{2} \qquad \forall n \in N, t \in T \tag{5}$$

$$\sum_m \left(y_{nmt} \sum_p \sum_r \left(Z_{rpt} \cdot \tau_p \cdot surg_{rm}\right) + y_{nm(t+1)} \sum_p \sum_r \left(Z_{rp(t+1)} \cdot \tau_p \cdot surg_{rm}\right)\right) + e_t + e_{t+1}$$

$$\le cap_t + cap_{t+1} + brk_t + ot \qquad \forall n \in N, t \in \{1,3,5,\ldots T-1\} \tag{6}$$

$$ut \cdot ra_{nt} - e_t - \sum_m \left(y_{nmt} \cdot \sum_p \sum_r \left(Z_{rpt} \cdot \tau_p \cdot surg_{rm}\right)\right) \le U_{nt}^{-} \qquad \forall n \in N, t \in T \tag{7}$$

$$\sum_m \left(y_{nmt} \cdot \sum_p \sum_r \left(Z_{rpt} \cdot \tau_p \cdot surg_{rm}\right)\right) + e_t - cap_t - brk_t \le U_{nt}^{+} \qquad \forall n \in N, t \in T \tag{8}$$

$$\sum_r \sum_t \left\lceil \frac{t}{2} \right\rceil \cdot Z_{rpt} - dd_p \le W_p \qquad \forall p \in P \tag{9}$$

$$\left(\frac{T}{2} + 1 - dd_p\right) \cdot \left(1 - \sum_r \sum_t Z_{rpt}\right) \le W_p \qquad \forall p \in P \tag{10}$$

$$U_{nt}^{+}, U_{nt}^{-}, W_p \geq 0 \qquad \forall n \in N, p \in P, t \in T \tag{11}$$

$$Z_{rpt} \in \{0,1\} \qquad \forall r \in R, p \in P, t \in T \tag{12}$$

The objective (1) minimizes the weighted total of under- and over-utilization costs and waiting time of patients after due dates. Due dates could be determined by the threshold that the hospital imposes on patient waiting time, or estimated by the surgeons. Constraints (2) ensure that each patient is assigned to at most one surgeon in the same specialty, only if the surgical group that the surgeon is in is assigned to that block. Constraints (3) indicate that each patient must be assigned to a surgeon that can be assigned to this patient, as sometimes surgeons bring their own patients to the hospital. Constraints (4) guarantee that each patient is assigned to at most one surgeon at one time. Constraints (5) and (6) ensure that the operating room is scheduled within the capacity and overtime limit in each individual block and each day, respectively. Constraints (7) define the under-utilization as the difference between the target usage and the total surgery time in a block, if the operating room is available and there is under-utilization. Constraints (8) define the over-utilization as the difference between the sum of all surgery time and the sum of capacity and lunch break of a block, if there is overtime. For a *scheduled* patient *p*, if scheduled after due date, the waiting time on the waiting list is defined in constraints (9) as the difference between the date that he/she is scheduled and the due date. For an *unscheduled* patient *p*, since he/she will be scheduled at least one day after the planning horizon $T/2$, the

waiting time is defined in constraints (10) as the difference between ($T/2+1$) and the due date. Constraints (11) and (12) are the integrality constraints.

In Phase 2 patients are sequenced in each room block.

Notation:

$n \in N$	index of OR
$b \in B$	index of beds in PACU
$p \in P$	index of patient
$w \in 2N$	index of room block
$k \in \{0,1\}$	index of stage, 0 if a patient is in OR, 1 if a patient is in PACU.
τ_p	estimated surgery time of patient p
υ_p	estimated recovery time of patient p
l	a very large number
α_2	relative weight factor of the objectives
tm	regular morning hours including lunch break
cap	capacity in each room

Decision variables:

OT_n	overtime in OR n
BM	total number of beds in PACU
MW_w	makespan in room block w
X_{pk}	start time of patient p in stage k
S_p	recovery time of patient p in OR

$OR_{pp'}$ 1 if patient p proceeds patient p' in the same OR, 0 otherwise

$PACU_{pp'}$ 1 if patient p proceeds patient p' in the same PACU bed, 0 otherwise

F_{bp} 1 if patient p is assigned to bed b, 0 otherwise

Note that we have a decision variable S_p to represent the recovery time of patients in OR. This is because of our assumption (4), which indicates patients have to recover in OR if there is no available bed in PACU. The result of Z_{rpt} in Phase 1.2 is input to Phase 2 and presented as z_{wp}, indicating if patient p is assigned to room block w.

MIP formulation of phase 2:

Min $\alpha_2 \sum_n OT_n + (1-\alpha_2)BM$ (1)

$$X_{p0} + \tau_p + S_p = X_{p1} \qquad \forall p \in P \tag{2}$$

$$z_{wp} \cdot (X_{p0} + Dur_{p0} + S_p) \le MW_w \qquad \forall p \in P, w \in 2N \tag{3}$$

$$TM + MW_{2n+1} - cap \le OT_n \qquad \forall n \in N \tag{4}$$

$$MW_{2n} + MW_{2n+1} - cap \le OT_n \qquad \forall n \in N \tag{5}$$

$$X_{p0} + \tau_{p0} + S_p \le X_{p'0} + l \cdot (3 - OR_{pp'} - z_{wp} - z_{wp'})$$

$$\forall w \in 2N, p \in P, p' \in P, p \ne p' \tag{6}$$

$$X_{p'0} + \tau_{p'} + S_{p'} \le X_{p0} + l \cdot (2 + OR_{pp'} - z_{wp} - z_{wp'})$$

$$\forall w \in 2N, p \in P, p' \in P, p \ne p' \tag{7}$$

$$X_{p1} + \upsilon_p - S_p \le X_{p'1} + l \cdot (3 - PACU_{pp'} - F_{bp} - F_{bp'})$$

$$\forall b \in B, p \in P, p' \in P, p \neq p' \tag{8}$$

$$X_{p'1} + \upsilon_{p'} - S_{p'} \leq X_{p1} + l \cdot \left(2 + PACU_{pp'} - F_{bp} - F_{bp'}\right)$$

$$\forall b \in B, p \in P, p' \in P, p \neq p' \tag{9}$$

$$b \cdot F_{bp} \leq BM \qquad \forall p \in P, b \in B \tag{10}$$

$$\sum_{b} F_{bp} \leq \left\lceil \frac{\tau_p - S_p}{\tau_p - S_p + 1} \right\rceil \qquad \forall p \in P \tag{11}$$

$$S_p \leq \upsilon_p \qquad \forall p \in P \tag{12}$$

$$X_{pk}, OT_n, S_p, MW_w, BM \geq 0 \qquad \forall n \in N, p \in P, w \in W \tag{13}$$

$$OR_{pp'}, PACU_{pp'}, F_{bp} \in \{0,1\} \qquad \forall b \in B, p \in P, p' \in P \tag{14}$$

The objective (1) minimizes the weighted total of overtime in all ORs and the largest number of beds used in PACU during the day, because as we indicated earlier, the staffing level in PACU is usually determined by the peak demand in hospitals. In this phase, constraints (2) ensure that each patient completes the operation and recovery in OR before transferred to PACU. Constraints (3) guarantee that all patients assigned to a room block finish their operations within the makespan of that room block. Constraints (4) and (5) define the overtime of an OR as the difference between the summation of makespan in the morning and the afternoon, and the daily capacity, if there is overtime in the morning; if not, it is defined as the difference between the summation of morning capacity including lunch break and the makespan in the afternoon, and the daily capacity. Constraints (6) and (7) indicate that an OR cannot have more than one patient scheduled at a time. Constraints (8) and (9) ensure that a bed in PACU not be occupied by more

than one patient at a time. Constraints (10) indicate that the assigned beds have to be less than the maximum number of beds in PACU. Constraints (11) indicate that all patients who are not fully recovered in OR must be scheduled in PACU and can only be assigned to one PACU bed. Constraints (12) guarantee that the recovery time of patients in OR no to exceed the estimated recovery time. Constraints (13) and (14) are the integrality constraints.

4.3.Heuristics and RKGA

Since phase 2 is done on a daily basis, we construct heuristics and RKGA for this phase. The complexity of doing an exhaustive search is first analyzed. Suppose we have 6 patients in each room, and 3 patients in each room block. Thus there are 16 room blocks (8 rooms with 2 time blocks each), the complexity of the patient sequencing problem is then $(3!)^{16} = 2.8 \times 10^{12}$. In general, the complexity is $O((p!)^n)$, where p is the number of patients in each room block, n is the number of room blocks.

Heuristic 1 – Johnson's rule

The first heuristic is to minimize the overtime without considering the number of beds in PACU ($\alpha_2 = 1$). Since this phase is similar to a two-step flow shop scheduling problem, we apply Johnson's rule for each OR, i.e. order all the patients and find the start and finish times in OR and PACU.

Heuristic 2 – Minimum beds

The second heuristic is to minimize the number of beds in PACU without considering the overtime ($\alpha_2 = 0$). We fix the number of beds to one to minimize

the objective. Then we order patients in each room block in increasing order of recovery time. Patients cannot leave the OR until they are fully recovered in the OR or the PACU bed is available.

Heuristic 3 – Modified Johnson's rule

The third heuristic aims to compromise to both objectives ($\alpha_2 = 0.5$). We first apply Johnson's rule to order patients as in Heuristic 1 and get the solution, then reduce the number of beds by half (if non-integral, take the ceiling). Fix the number of beds all through the day. Keep the order of patients as in Johnson's rule, but similar to Heuristic 2, patients cannot leave the OR until they are fully recovered in the OR or the PACU bed is available.

Random Keys Genetic Algorithm

Introduced by Holland (1975), Genetic Algorithm (GA) is an adaptation procedure based on the mechanics of natural genetics and natural selection. GA efficiently searches the solution space globally by combining the existing solutions to form new ones. We refer to (Davis, 1991) and (Goldberg, 1989) for a detailed introduction to genetic algorithms.

GA starts by initializing a population, of which each individual "chromosome" represents a solution of the problem in the form of a string structure. Then a fitness value is calculated to assess the relative quality of each individual. The optimization process of GA takes advantage of three GA operators: selection, crossover, and mutation. The selection operator uses the fitness value to adjust the survival probability of each individual in the population. The probabilities are used to randomly select survivors to generate offsprings. The

crossover operator combines pairs of individuals in the current population (parents). The mutation operator chooses a random position in a chromosome, and changes the value to a new randomly selected value.

A common problem for combinatorial applications of genetic algorithms is that some operations may create infeasible solutions. Attentions of researchers have been attracted to fix this problem in different ways, the most commonly used ones of which are to "repair" the algorithm repeatedly after a generation to recreate only feasible solutions. However, the repair is computational expensive and may cause convergence (Michalewicz, 2000; Haral et al., 2006). Bean (1994) has introduced an alternative method to encode problem solutions using random numbers called RKGA, which is known as a better alternative for this type of GA applications. RKGA differs from traditional GA mostly in the solution representation. Specifically, a random number encoding structure is used in the chromosomal representation to avoid creating infeasible chromosomes during traditional GA crossover. In our study, the chromosome is represented in the form of ROOM_BLOCK.KEY. For example, the representation of a chromosome in the patient sequencing problem in two room blocks would be in the following form, as an instance: (2.93854, 2.75581, 1.28560, 2.00645, 1.65938). Each number represents a patient. The part of the number to the left of the decimal is used to assign *room blocks* and the part to the right is used to assign the *sequence*. In the example chromosome, the first, second and fourth patient would go to room block 2, and the fourth patient is scheduled first because it has the smallest key,

followed by the second and the first one. Similarly, the third and the fifth patient would go to room block 1, and the third one is before the fifth one.

4.4.Interactions between two phases

As we stated in section 2, most of the literature has decomposed the problem into two phases. However, both these steps interact with each other in reality. The allocation of capacity may affect the performance of daily scheduling. Therefore, in this study we also consider treating both Phase 1.2 and 2 simultaneously and compare with decomposed solutions. We do not consider Phase 1.1 because this phase is a higher level planning of capacity, in which decisions are made upon how much capacity to allocate to each surgical group. This is done for a much longer planning horizon (sometimes longer than a year) and patients and individual surgeons are not involved. Thus, we start by developing an MIP model that includes both the planning over a short time horizon (one or two weeks) and the scheduling in each single day. The model can be found in Appendix A. Then, we use the same data for both situations to see how much loss of optimality there would be from the decomposition.

5. Computational results

5.1.Input data

Our data is from an outpatient clinic of a major healthcare provider. There are eight operating rooms, four clinical specialties with a total of 36 surgeons. The average case duration and number of surgeons in each specialty is shown in Table 10. When blocks are fixed for surgical groups, mean surgery durations are typically used to determine whether the cases fit in the block (2007).

By summing up the surgery time and number of patients for all consecutive two weeks, from August 2005 to July 2006, the maximum, mean and minimum bi-weekly surgery demand are identified and present in Table 11.

Table 10

Data analysis by specialty

Specialty	*Average case duration (min)*	*Number of surgeons*
Pain clinic	26	11
Urology	60	8
Ophthalmology	51	12
Oral Maxillofacial	40	5

Table 11

Bi-weekly demand analysis

	Dates	*Bi-weekly total surgery time*	*Bi-weekly total number of patients*
Maximum	06/05/2006 - 06/16/2006	18350	451
Mean	02/13/2006 - 02/24/2006	15781	377
Minimum	12/26/2005 - 01/06/2006	9628	264

5.2.Results of Phase 1

We conduct experimentation for each of the MIP formulations. These formulations are modeled with C++ and CPLEX version 11.0 is used to solve the problem instances. The experiments were run on a 2.66GHz PC with 4GB RAM.

The objective value of optimal solutions of Phase 1.1 is simply the number of unsatisfied surgeon preferences. In the data there is no record of surgeon preferences. We suppose all surgeons choose 30% of the capacity to be their preferred time within their available time. Surgeons can be assigned available but not preferred blocks, although that would cause the increase in objective value.

We have three levels in demand (maximum, mean and minimum). To analyze the impact of the number of surgical groups, we use four levels (1, 2, 3 and 4 groups in each specialty). Optimal solutions of Phase 1.1, which is the number of unsatisfied preference in each scenario, is listed in Table 12.

Table 12

Computation results of Phase 1.1

Number of groups	*1*	*2*	*3*	*4*
Min Demand	9	11	12	15
Mean Demand	13	15	18	24
Max Demand	16	19	22	33

The optimal solutions of Phase 1.1 for different demand volumes are input to Phase 1.2. In this phase, we have three levels in demand and four levels of surgical groups as in Phase 1.1, and five different values of weight factor α_1 (0.1, 0.3, 0.5, 0.7, 0.9). The results of Phase 1.2 are presented in Figure 6. Different shapes of dots in the figures represent the results with different number of groups in each specialty as indicated in the top right legend.

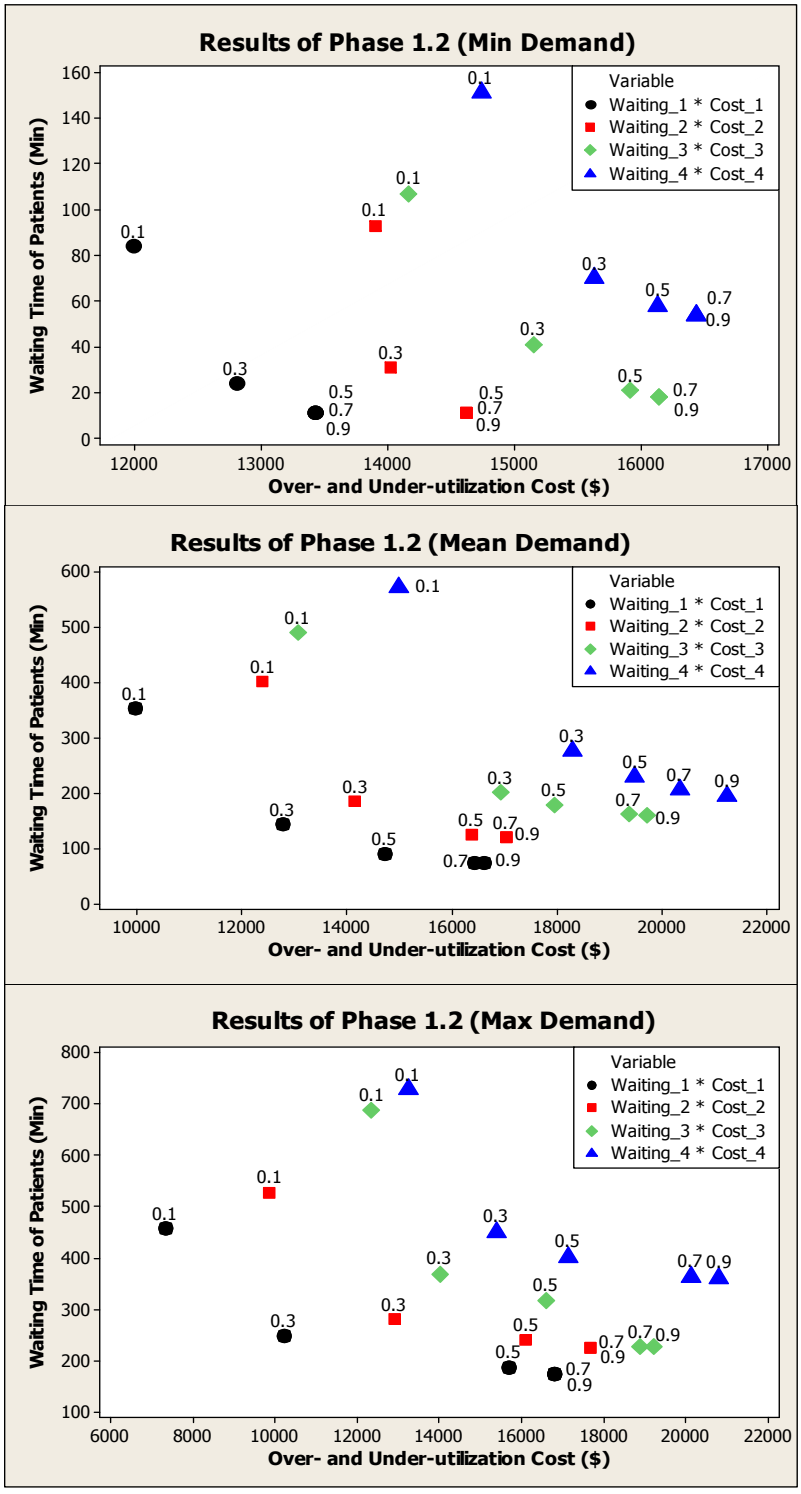


Figure 6. Results of Phase 1.2

As can be seen from the results in both Phase 1.1 and Phase 1.2, the performance of more groups in each specialty is dominated by that of fewer groups. This is as expected because having more surgeons in a group brings more flexibility. The result also supports the findings in several previous studies that the block scheduling approach is preferred over the open scheduling approach because the former yields a better utilization, as mentioned in section 3.2.

Then we choose the data with mean demand as a representative to analyze the weighted total of two objectives to find the relationship between the increment in the number of groups and the objective value. In Figure 7, the connected lines represent the change in objective value with different weight factors while increasing the number of groups. As we can see from the figure, objective value increases nonlinearly as the number of groups increases.

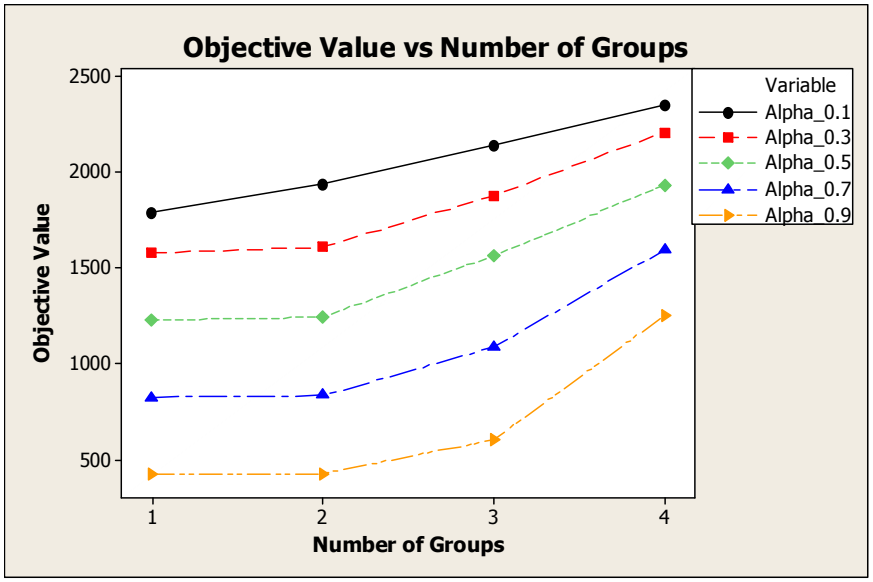


Figure 7. Impact of changing number of groups in weighted sum of multi-objectives

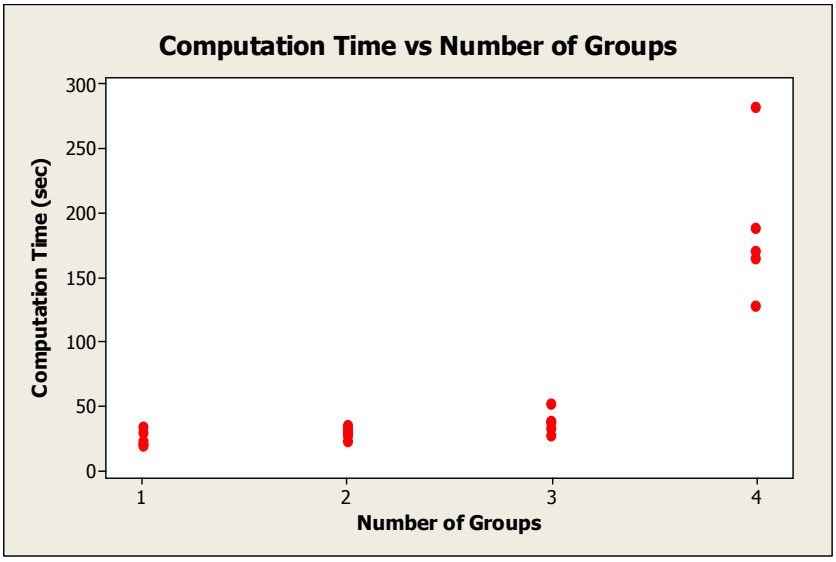


Figure 8. Impact of changing number of groups in computation time

Figure 8 shows the change in computation time with different numbers of groups. As is shown in the figure, when the number of groups goes up from 1 to 3, not much increase in computation time is observed, but at 4 groups there is a dramatic increase. Thus from our results we conclude that there is a nonlinear increasing trend in computation time when to the number of groups goes up.

5.3.Results of Phase 2

One day is picked randomly and we solve daily scheduling problem using MIP, RKGA and all the heuristics we proposed. The comparison of the results is can be found in Figure 9 and Table 13 (GA Population size: 1000. Number of generation: 500).

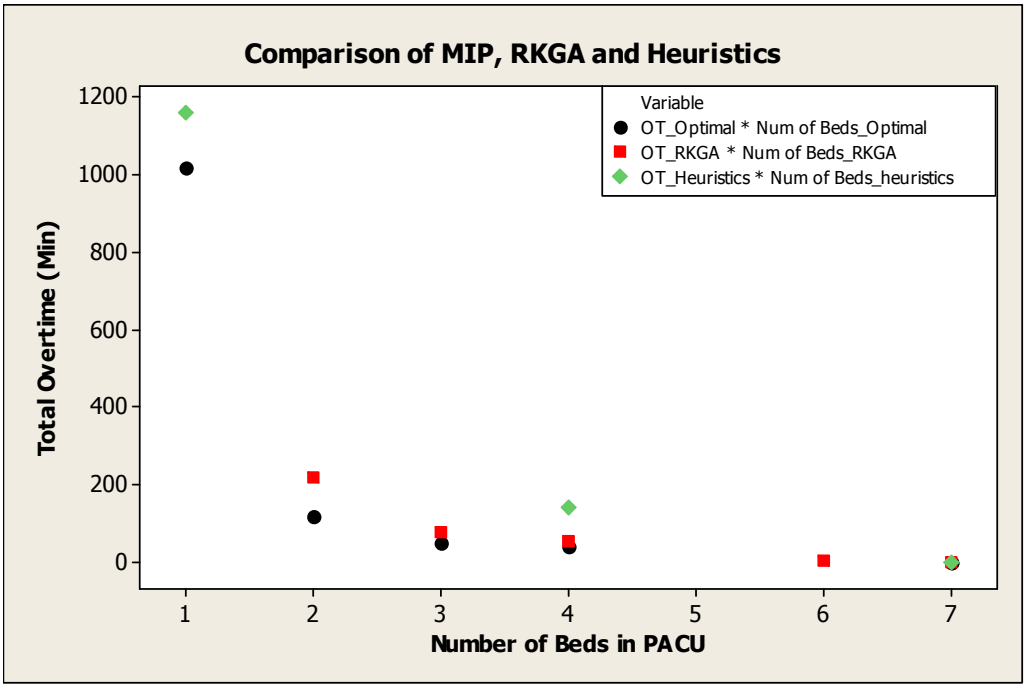


Figure 9. Comparison of MIP, RKGA and heuristics

As can be seen from the results, RKGA perform very close to optimum while using much less computation time than the MIP. The heuristics are simple to implement, consuming even less time, but Heuristic 1 and 2 are lacking the compromise between the two objectives. Figure 9 illustrates the trade-off between the objectives. Managers can make decisions on the schedule based on their own criteria.

Table 13

Comparison of the results with MIP, RKGA and Heuristics

α_2	*MIP*			*RKGA*			*Heuristics*		
	Over-time	*Number of Beds*	*Comp. Time (sec)*	*Over-time*	*Number of Beds*	*Comp. Time (sec)*	*Over-time*	*Number of Beds*	*Comp. Time (sec)*
0	1020	1	7258	217	2	162	1159	1	38
0.2	119	2	19320	76	3	289	143	4	71
0.4	48	3	27256	56	4	125			
0.6	38	4	20667	56	4	311			
0.8	0	7	34153	2	6	284			
1	0	7	5114	0	7	192	0	7	23

5.4.Analysis of interactions between two phases

In the combined model in Appendix A, room blocks are used in both phases instead of time blocks to standardize the time unit. The cost of under-utilization, overtime and staffing in PACU is combined to one objective, and waiting time of patients is the other. We fix the relative weight factor to 0.5. In the decomposed model, we use the same value of weight factor. Since both Phase 1.2 and Phase 2 aim to minimize the cost of overtime, the optimal objective value will only be taken from Phase 2. Due to the increase in complexity, the number of days in a planning horizon is limited to one week. The experimental design and results can be found in Table 14.

Table 14

Analysis of interactions between two phases

Number of days	*Number of rooms*	*Gap to optimum by decomposition*			*Computation time*		
					Average time (sec)		*Difference*
		Min Demand	*Mean Demand*	*Max Demand*	*Optimum*	*Decom-posed*	
3	2	1.29%	1.84%	2.89%	42459	77	99.82%
	4	2.33%	3.38%	4.37%	73681	88	99.88%
	6	3.76%	4.96%	5.91%	89878	135	99.85%
	8	5.17%	6.23%	7.10%	116763	146	99.87%
5	2	2.04%	2.71%	3.82%	86642	91	99.89%
	4	3.28%	4.50%	5.60%	128939	126	99.90%
	6	4.85%	6.21%	7.73%	169086	168	99.90%
	8	7.22%	9.06%	10.51%	193771	219	99.89%

Overall, the solution obtained from the decomposed model is close to optimal (1%-11% gap). At the same time, the computational time is greatly reduced by around 99% (from several hours to less than an hour) through decomposing the two phases. The results also show the impact of interaction increases as 1) the planning horizon increases; 2) the size of operating department increases; 3) the demand increases. The impact of number of rooms is slightly greater than the number of days. Consequently, in practice if there are many rooms in the facility or the demand is very high, to decrease the effect of decomposition, hospitals could choose a shorter planning horizon for Phase 1.2.

6. Conclusions

In this chapter we have introduced a modeling approach to OR planning and scheduling. The problem is modeled in two phases with MIP. We consider

multiple criteria while evaluating the performance of the planning and scheduling, including OR utilization and overtime, surgeons' preference, patient waiting time and patient flow from OR to PACU. The exact solution from Phase 1 illustrates the trade-offs between operational objectives and patient/surgeon satisfaction objectives. Also it shows that fewer number of surgeon group yields better performances. In the second phase, we first obtain the optimal solution from MIP. Due to the complexity and computation time considerations, three heuristics and RKGA are developed, from which close-to-optimal solutions can be derived much more efficiently. The impact of decomposing the problem is found to be ~11% loss in the optimality, which is tolerable considering the 90% saving in computation time.

In the study we assumed that all procedure times are deterministic. A simulation model would be a useful extension to the study. The optimal solution from the MIP model can be tested in the simulation model that captures some of the randomness of the processes (for instance, surgery time, demand, and arrival time).

CHAPTER 4

A MULTI-OBJECTIVE SIMULATION OPTIMIZATION APPROACH TO OPERATING ROOM SCHEDULING

1. Introduction

The surgery scheduling problem involves several conflicting objectives, such as patient satisfaction and operational cost. Improving one objective may depreciate the performance of one or more other objectives. Traditional approaches for solving multi-objective optimization problems try to scalarize the multiple objectives into a single objective and change the problem formulation into a single objective optimization problem in which only one global optimal point is desired. However, there are several drawbacks to scalarize objectives, such that the priority vector is playing a key role in the final solution, and some alternative solutions may not be available to decision makers without changing the priority vector. Although some optimization techniques, such as goal programming, genetic algorithms (GA), and simulated annealing have been used to deal with multiple objectives, they often fail to capture the uncertainties in health care practice.

DES is a powerful tool in evaluating complex health care systems and answering "what if" questions. There have been extensive studies on using DES to study health care operations (Dexter et al., 1999b; Everett, 2002). It allows hospital managers to include most of the randomness in reality. However, practical questions are often seeking optimum values for the decision variables and thus exploratory process for optimal solutions is needed.

Simulation optimization is the process of finding the best values of some decision variables for a system where the performance is evaluated through simulation (Fu, 2002). It conquers the difficulties in optimization to incorporate randomness and guides the simulation to find the optimal solution efficiently. Due to the uncertainty nature in the health care industry, there have been some efforts on applying simulation optimization in health care (Angelis et al., 2003; Ahmed & Alkhamis, 2009; Baesler et al., 2001) while no previous literature is found to apply simulation optimization in OR scheduling to our best knowledge. In this chapter, we use simulation optimization to model and solve the surgery scheduling problem. By combining RKGA and NSGA-II as the optimization algorithm in multi-criteria simulation optimization, it can also be applied in general scheduling problems.

In addition, we would like to investigate answers to the following important questions from the managerial perspective:

1) What is the optimal length of time block for each case?
2) How much impact does patient no-show have on the scheduling performance?
3) How much impact does the downstream resource have on the scheduling?

The rest of the paper is organized as follows. The next section gives an overview of previous literature related to our study. In section 3, the problem is described and modeled as mixed integer programs (MIP). Our simulation optimization methodology is illustrated in section 4, followed by experimental results in

section 5. In section 6 the managerial questions are analyzed. Finally, we discuss our conclusions and future research directions.

2. Literature review

2.1.Operational research in OR scheduling

Operating room scheduling problems have gained much attention in the operational research area and have been extensively studied recently due to the increased importance of providing health services efficiently and effectively. Cardoen et al. (2010) provides a review of recent operational research literature on operating room planning and scheduling. One of the major problems associated with the development of accurate OR scheduling is the uncertainty inherent to surgery services. Deterministic scheduling approaches ignore such uncertainty or variability, which is essential for solving realistic problems. Stochastic approaches try to incorporate uncertainties related to surgery durations and patient arrivals (Cardoen et al., 2010; Erdogan & Denton, 2009). However, many other aspects of uncertainty in reality, including availability of downstream resources, patient no-shows and accommodation of add-on cases that arise on short notice, are still open in existing stochastic optimization literature on surgery scheduling (Denton et al., 2007; Denton et al., 2009).

2.2.The use of simulation optimization in health care

There have been several efforts in developing simulation optimization models for solving problems in healthcare management in the last decade, though none has been found in surgical scheduling. Angelis et al. (2003) use simulation, estimation of target function and optimization interactively to assign servers and

facilities to different services in a health care. Ahmed and Alkhamis (2009) design a decision support system for the operation of an emergency department that uses simulation optimization to determine the optimal number of staff to maximize patient throughput and to reduce patient time in the system subject to budget constraints. Baesler and Sepulveda first introduce an approach by integrating GA with goal programming (2000), and then apply their methodology to design a cancer treatment facility (2001). They consider patients' waiting time, closing time, and nurse and chair utilization as performance measures.

2.3.Multi-objective simulation optimization

Most of the applications of simulation optimization have been single objective problem. In the literature there are limited attempts to multi-objective simulation optimization problems (Table 15). The majority of them are focused on response surface methodology and interactive procedures. The major drawbacks are local optimality and lack of automated direct search.

There have been a few papers considering operation scheduling problems using simulation optimization. Almeida et al. (2003) introduce a simulation-based approach for multi-objective optimization of operation scheduling in a petroleum refinery, which is based on GA combined with a multi-objective energy minimizing method. Allaoui and Artiba (2004) use a combined method of simulated annealing and dispatch rules for flow shop scheduling problems. Gupta and Sivakumar (2002) propose an approach based on compromise programming for operation scheduling in semiconductor manufacturing and apply the method to find a Pareto optimal solution of a NP-hard problem.

Table 15

Summary of multi-objective simulation optimization literature

Author (Year)	*Methodology*	*Application*
Mollaghasemi et al. (1991)	Integrate gradient search and multiple attribute value function	
Mollaghasemi & Evans (1994)	STEP method (to minimize the maximum deviation of objectives from the ideal solution using relative weight of deviations)	A job shop model
Teleb & Azadivar (1994)	Interactive approach	
Boyle & Shin (1996)	Interactive approach	
Beasler & Sepulveda (2000)	Integrate GA, goal programming and	Cancer treatment facility design (Beasler & Sepulveda, 2001)
Joines et al. (2002)	Modified NSGA-II	Supply chain optimization
Gupta & Sivakumar (2002)	Compromise programming	Scheduling in semiconductor
Almeida et al. (2003)	GA combined with multi-objective energy minimizing method	Scheduling in petroleum refinery
Allaoui & Artiba (2004)	Simulated annealing combined with dispatch rules	Flow shop scheduling
Eskandari et al. (2005)	Integrate stochastic nondomination-based multi-objective optimization technique and GA	
Willis & Jones (2008)	Heuristic search algorithm and database technologies	
Zsakerifar et al. (2009)	Kriging metamodeling	(S,s) inventory system

2.4.Original contributions of this research

This study develops a modeling framework using simulation optimization to assist the OR scheduling in hospitals, serving as an alternative which conquers the difficulties in pure simulation and optimization. We take into account of uncertainty in practice, including actual start time and duration of surgeries, downstream resources and patient no-shows.

Multiple objectives are considered, including patients' waiting time and operational cost composed of overtime and under-utilization cost, the staffing cost in Post Anesthesia Care Unit (PACU), and the fixed cost of opening OR. RKGA and NSGA-II are combined as the optimization algorithm in multi-criteria simulation optimization for the first time. Pareto optimal solutions are compared and shown to be outperforming single objective simulation optimization and pure GA.

3. OR scheduling problem formulation

In this study, we investigate the surgery scheduling problem which consists of multiple OR and a set of patients, under uncertainty. The objective is to minimize patient waiting and operational cost. Each patient goes through two stages: surgery in OR and recovery in PACU, both having stochastic durations. There is a possibility that some patients do not show up for the surgery.

Block scheduling is used, with which a block of time (usually one-half or a full day) is allocated to one surgeon. There is a lunch break in the middle of the day. Patients are then assigned to blocks and reserved a certain period. The length of the period is usually determined by the distribution of the particular type of patients. The planning horizon can vary from one to several days. We assume all

patients are independent and available to be scheduled at time 0, and if showing up on the day of surgery, patients arrive at the beginning of the period. Upon arriving, if the previous patient is not finished in the OR, a patient has to wait until the previous patient finishes. PACU resource is assumed to be shared by patients from all OR.

The problem is formulated as a mixed integer program as follows.

Deterministic Parameters

$r \in R$	index of room
$t \in T$	index of day
$s \in S$	index of specialty
$i \in I$	index of room block
$p \in P$	index of patient
brk_i	length of the lunch break after room block i
ra_i	1 if room block i is available to schedule cases, 0 otherwise
rs_{rs}	1 if patients in specialty s can be assigned to room r, 0 otherwise
ot	maximum possible overtime per room per day
α_1	relative weight factor of objectives in model 1.2
co, cu	cost of overtime and under-utilization, respectively
cr	fixed cost of opening an OR
cb	cost of staffing in PACU
e_i	capacity reserved for emergency patients in block i (*from historical data*)

ε_p estimated surgery time of patient p

pts_{ps} 1 if patient p is in specialty s, 0 otherwise

ptb_{pi} 1 if patient p has to be assigned to room block i, 0 otherwise

cap_i total capacity in room block i

λ relative weight factor of the two objectives

Random Parameters

τ_p actual surgery time of patient p

υ_p actual recovery time of patient p

ns_p 1 if patient p shows up, 0 otherwise

Decision variables

AS_{pi} 1 if patient p is assigned room block i, 0 otherwise

$OR_{pp'}$ 1 if patient p proceeds patient p' in the same OR, 0 otherwise

$PACU_{pp'}$ 1 if patient p proceeds patient p' in the same PACU bed, 0 otherwise

F_{bpt} 1 if patient p is assigned to bed b on day t, 0 otherwise

P_i 1 if room block i is open, 0 otherwise

S_p recovery time of patient p in OR

Resultant variables

OT_i overtime in room block i

UT_i utilization in room block i

BM_t number of beds in PACU used on day t

$X1_{pi}$ start time of patient p in OR in room block i

$X2_{pt}$ start time of patient p in PACU on day t

W_p waiting time of patient p after their scheduled time

ARR_p scheduled arrival time of patient p

MIP Formulation

$$\textbf{Min } \lambda \cdot \left(\sum_i \left(co \cdot OT_i + cu \cdot UT_i + cr \cdot P_i \right) + c_b \cdot \sum_t BM_t \right) + (1-\lambda) \cdot \sum_p W_p \tag{1}$$

$$AS_{pi} \le ra_i \cdot rs_{\lceil i/2T \rceil, s} \cdot P_i \quad \forall p \in P, i \in I \tag{2}$$

$$\sum_{i=2Nt}^{2N(t+1)} X1_{pi} + ns_p \left(\tau_p + S_p \right) = X2_{pt} \quad \forall p \in P, t \in T \tag{3}$$

$$\sum_p \left(AS_{pi} \cdot \varepsilon_p \right) + e_i \le cap_i + brk_i + \frac{ot}{2} \quad \forall i \in I \tag{4}$$

$$\sum_p \left(AS_{pi} \cdot \varepsilon_p + AS_{p(i+1)} \cdot \varepsilon_p \right) + e_i + e_{i+1} \le cap_i + cap_{i+1} + brk_i + brk_{i+1} + ot$$

$$\forall i \in \{1,3,...I-1\} \tag{5}$$

$$ns_p \cdot AS_{pi} \left(X1_p + \tau_p \right) \le OT_i \quad \forall i \in I \tag{6}$$

$$ut \cdot ra_i \cdot P_i - \sum_p \left(ns_p \cdot AS_{pi} \cdot \tau_p \right) - e_i - OT_i \le UT_i \quad \forall i \in I \tag{7}$$

$$ARR_p = \sum_{p'} \left(OR_{pp'} \cdot \varepsilon_p \cdot \left(\sum_i AS_{pi} \right) \right) \quad \forall p \in P \tag{8}$$

$$ns_p\left(\sum_i X1_{pi} - ARR_p\right) \le W_p \qquad \forall p \in P \tag{9}$$

$$X1_{pi} + ns_p \cdot \tau_p \le X1_{p'i} + l \cdot \left(3 - OR_{pp'} - \sum_i AS_{pi} - \sum_i AS_{p'i}\right)$$

$$\forall i \in I, p \in P, p' \in P, p \ne p' \tag{10}$$

$$X1_{p'i} + ns_p \cdot \tau_{p'} \le X1_{pi} + l \cdot \left(2 + OR_{pp'} - \sum_i A_{pi} - \sum_i A_{p'i}\right)$$

$$\forall i \in I, p \in P, p' \in P, p \ne p' \tag{11}$$

$$X2_{pt} + ns_p \cdot \upsilon_p \le X2_{p't} + l \cdot \left(3 - PACU_{pp'} - F_{bpt} - F_{bp't}\right)$$

$$\forall b \in B, t \in T, p \in P, p' \in P, p \ne p' \tag{12}$$

$$X2_{p't} + ns_p \cdot \upsilon_{p'} \le X2_{pt} + l \cdot \left(2 + PACU_{pp'} - F_{bpt} - F_{bp't}\right)$$

$$\forall b \in B, t \in T, p \in P, p' \in P, p \ne p' \tag{13}$$

$$b \cdot ns_p \cdot F_{bpt} \le BM_t \qquad \forall b \in B, t \in T, p \in P \tag{14}$$

$$\sum_b \sum_t ns_p \cdot F_{bpt} \le \left\lceil \frac{\tau_p - S_p}{\tau_p - S_p + 1} \right\rceil \qquad \forall p \in P \tag{15}$$

$$ns_p \cdot S_p \le \upsilon_p \qquad \forall p \in P \tag{16}$$

$$X1_{pi}, X2_p, W_p, ARR_p, S_p, BM_t, OU_i, UT_i \ge 0 \tag{17}$$

$$AS_{pi}, F_{bpt}, P_i, OR_{pp'}, PACU_{pp'} \in \{0,1\} \tag{18}$$

The objective (1) is to minimize the weighted total of cost and patients' waiting time. The cost is composed of the cost of overtime and under-utilization in the OR, the staffing cost in PACU on each day, and the fixed cost of opening OR. Constraints (2) ensure each patient is assigned to at most one room block that

is open, in the same specialty, and they can be assigned to that block. Constraints (3) ensure each patient completes all operations in OR before transferred to PACU. Constraints (4) and (5) guarantee that each OR block is scheduled within the capacity and overtime constraint. Constraints (6) and (7) define overtime and under-utilization, respectively. Constraints (8) define the scheduled start time of each patient. Constraints (9) indicate that the waiting time is the difference between the actual start time and the scheduled start time of a patient, if there is waiting for that patient. Constraints (10) and (11) ensure that OR cannot be occupied by more than one patient at a time. Constraints (12) and (13) guarantee that a bed in PACU not occupied by more than one patient at a time. Constraints (14) ensure the index of assigned beds is less than the maximum number of beds in PACU. Constraints (15) indicate that all patients who are not fully recovered in OR must be scheduled in PACU and can only be assigned to one PACU bed. Constraints (16) indicate that the recovery time in OR cannot exceed the actual recovery time needed for all patients

4. RK-NSGA-II based simulation optimization methodology

GA were introduced by Holland (1975) as a methodology to adaptively search for solutions to complex problems based on the mechanics of natural genetics and natural selection. The procedure involves representing solutions as “chromosomes” and generating new population of chromosomes through randomly choosing and changing chromosomes. In this study, the main optimization routine that we use for searching schedules is developed based an RKGA implementation of NSGA-II, which we call “RK-NSGA-II”. The

chromosomes are represented using the random number encoding structure of RKGA, and new populations are generated by the operators of NSGA-II. We will explain these two different types of GA in detail as follows.

4.1.RKGA

GA chromosomes are usually strings of numbers that represent the solution to the problem or can be decoded to represent the solutions. As an important operator of GA, crossover can cause infeasibility when applying GA to scheduling problems (Haral et al., 2006). Introduced by Bean (1994), RKGA uses random number encoding structure in the chromosomal encoding to avoid creating infeasible chromosomes during traditional GA crossover. For the surgery scheduling problem, if using a *p*-dimensional vector representing the order of *p* patients to for chromosomes, by applying crossover, two types of infeasibility may be created: (1) patients may be assigned to the room that may be constrained to their specialties; (2) some patients may be repeated or omitted. Using RKGA, a 2-dimensional vector consisting of two random numbers (keys) for each patient can avoid such problems. Two keys are generated randomly from 0 to 1. The first key decides which room block the patient would be assigned. The second key decides the sequence of patients in each room.

For example, the chromosome of one patient contains the following key: (0.3245, 0.1287). Assume the patient can be assigned to four room blocks (3, 4, 7, 8). Dividing 1 into four equal intervals, 0.3245 would fall in the second interval from 0.25 to 0.5. Thus the patient would go to room block 4. After all patients are

assigned to room blocks, they are sequenced according to the increasing order of the sequence key.

4.2.NSGA-II

NSGA-II was first introduced by Deb and Goel (2002). We adopt NSGA-II as our GA operator because it outperforms over other multi-objective GA in generating Pareto frontier (2006). In the NSGA-II evolutionary process, individuals are first selected from the current generation to be parents based on the fitness, which is determined by a ranking process for a Pareto-based multi-objective GA. The rank is determined by its Pareto dominance in the current population. To maintain a good spread of solution set, crowding distance is calculated to estimate the density of the individuals surrounding a particular individual in the population. It is done for a solution point by calculating the average distance of two points on either side of the point along each of the objectives. The logic of one generation of NSGA-II can be found in Figure 10.

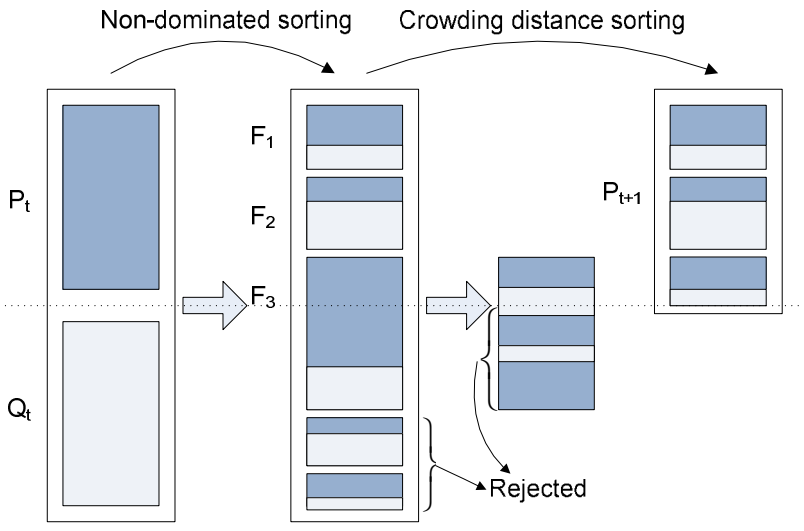


Figure 10. Main loop of NSGA-II

4.3.Modeling framework

The main logic of our approach is present in Figure 11. First, the information of patients is input and a set of configurations and the random keys are generated to form the initial population of GA. These individuals are then translated to surgery schedules which consist of the assignment and sequence of all patients. A simulation model is run for each configuration and the output is recorded. The values are ranked according to dominance and thus the nondomination frontiers can be found. Crowding distance is calculated to distinguish individuals that have the same rank. The Pareto optimal set is then updated and checked with stopping criteria. If the stopping criteria are not satisfied, the traditional GA selection, crossover and mutation are performed and the new population generated is repeating this process from the beginning. On the other hand, if not satisfying the stopping criteria, the current Pareto optimal set is the final result.

Elitist is guaranteed by the flow of the NSGA-II algorithm, i.e. the first front (which is the Pareto optimal set of a generation) is always kept in the next generation. A stopping criterion is adopted based on the *convergence speed* towards the Pareto optimal curve. If in a pre-specified number of consecutive generations, no considerable improvement is found in the quality of the Pareto optimal curve, the algorithm is stopped. Alternatively, the algorithm could be stopped after a specific number of generations.

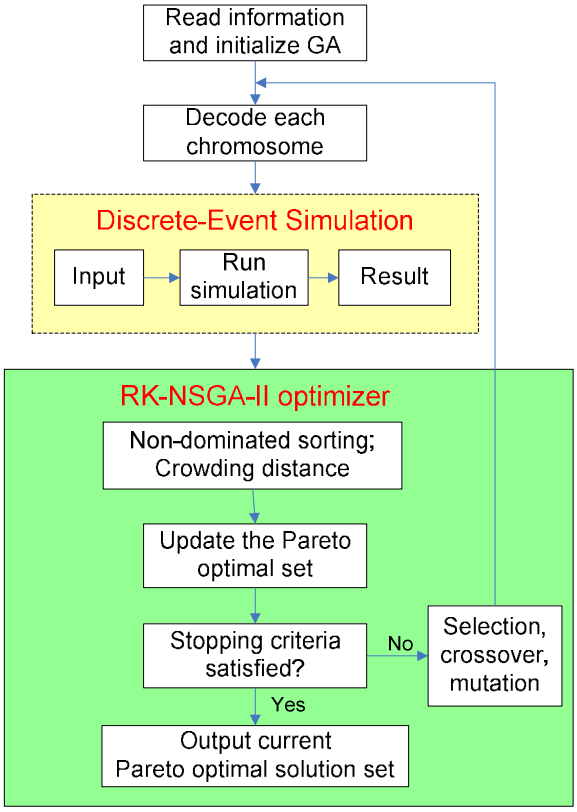


Figure 11. Simulation optimization framework

5. Computational experiments

5.1.Data description

The data in this study are provided by an outpatient clinic of a major health care provider in the US. A sample of 10570 surgeries from August, 2005 to February, 2007 is used. We categorize patients by specialty and allocate time accordingly, as there is statistical difference with 99% confidence among the surgery times. All surgery and recovery times follow Weibull distribution. The mean, 65^{th} percentile, 75^{th} percentile and 85^{th} percentile of each specialty are shown in Table 16. We use planning horizon of one week in this study, which can

be adjusted in practice. The maximum, mean and minimum number of patients per week in the sample data is present in Table 17.

Table 16

Surgery duration (min)

Specialty	*Mean*	*65th percentile*	*75th percentile*	*85th percentile*
Anesthesiology	27	26	30	36
Urology	60	60	72	97
Ophthalmology	51	48	56	71
Oral & Maxillo Surg	40	43	49	58

Table 17

Weekly demand

	Dates	*Number of patients*
Maximum	06/12/2006 - 06/16/2006	228
Mean	08/29/2006 - 09/02/2006	196
Minimum	11/21/2005 - 11/25/2005	159

5.2.Implementation of the simulation-optimization methodology

The simulation optimization model is implemented in C++ and run on a PC with a 2.66GHz processor with 4GB of RAM. We first investigate the convergence of the algorithm.

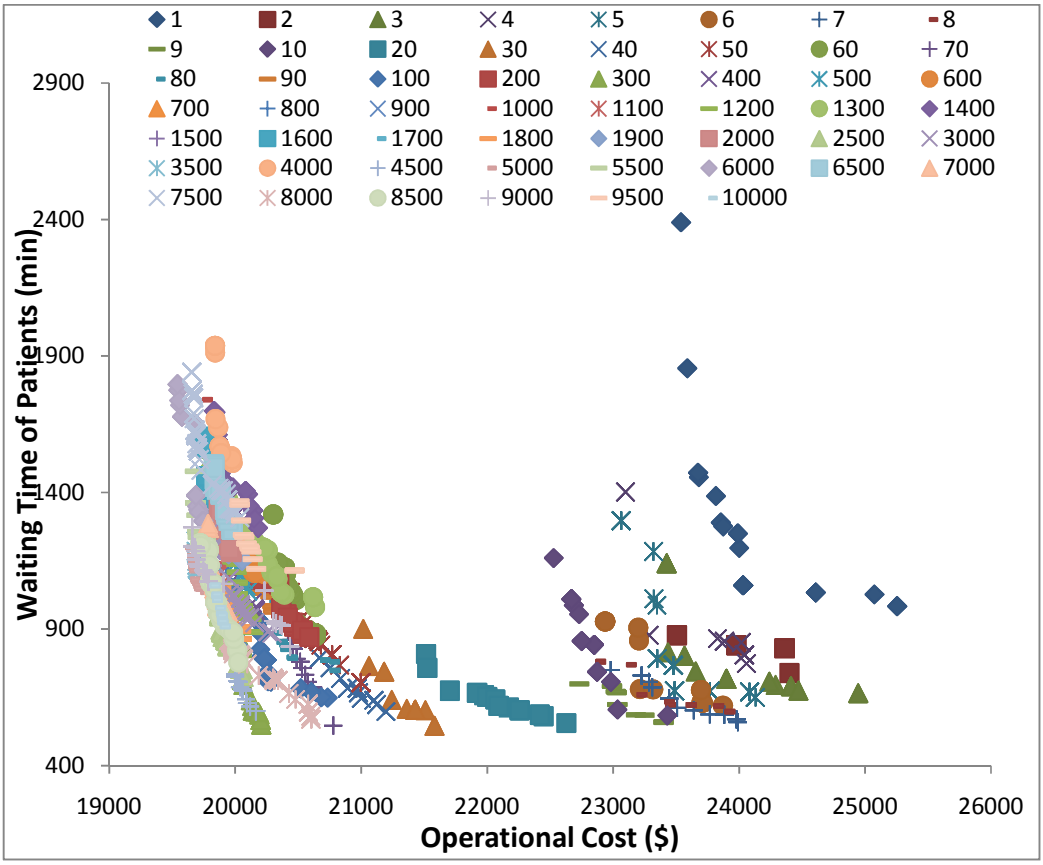


Figure 12. Efficient Frontier in 1 – 10000 generations (GA population size: 2000)

Surgery and recovery durations are randomly generated in the simulation module according to the distribution of each specialty. The basic structure as shown in Figure 12 is seen in all experiments with different demand pattern and different length of duration allocated to each patient. The efficient frontier is improved substantially while increasing the number of generations from 1 to 1000. Starting from 1000 generations, the variation in efficient frontier between every 500 generations is much smaller, and the movement of efficient frontier is random rather than converging to the ideal point (Figure 13).

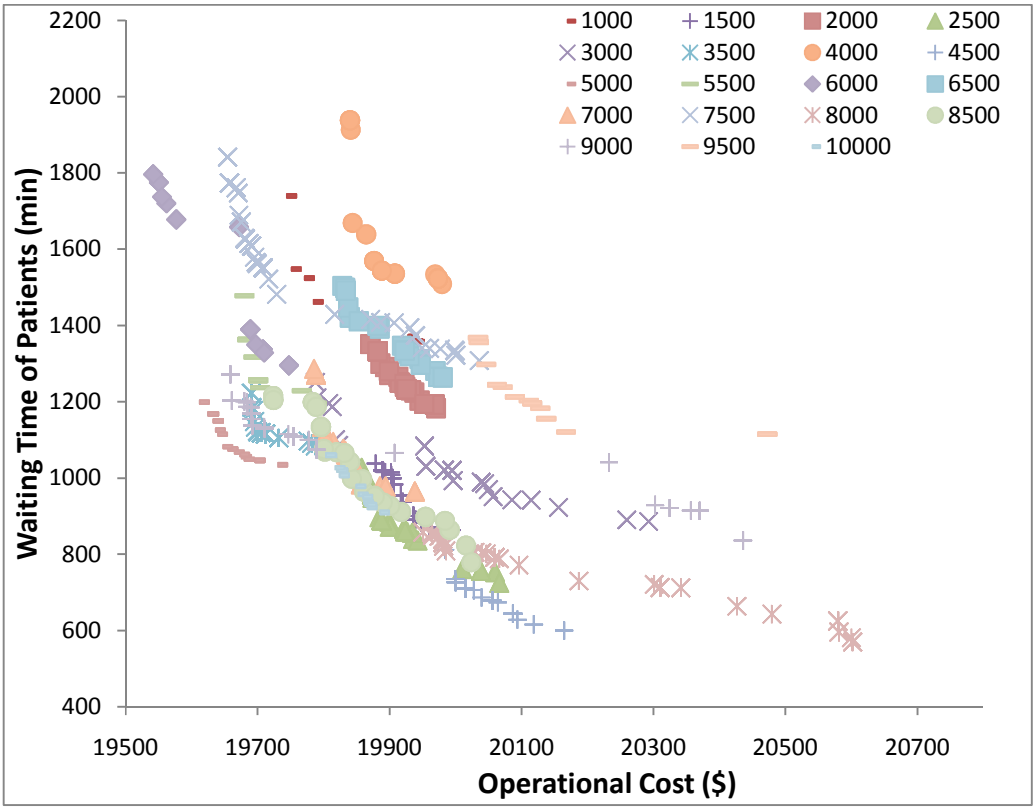


Figure 13. Efficient Frontier in 1000 – 10000 generations

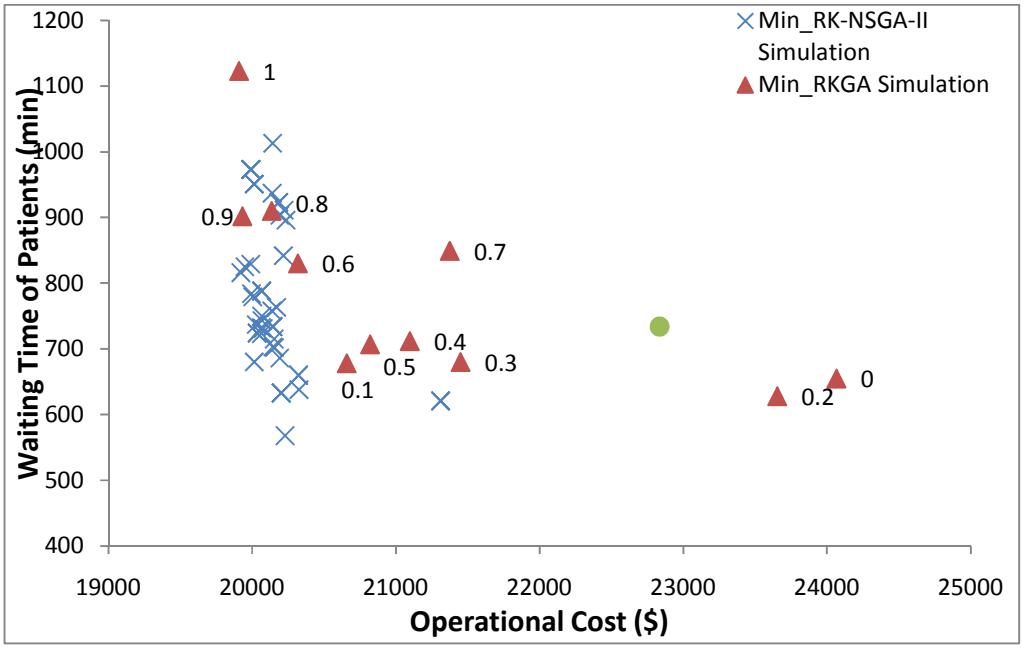
The random movement can be interpreted that the variation is caused by the randomness in simulation, not the number of generations. Thus we decide to choose 1000 as the minimum generations. For the stopping criterion, as stated in section 4, the *convergence speed*, if the improvement in both objectives is less than 10% in 50 consecutive generations, the algorithm is stopped. Alternatively, the algorithm could be stopped after a sufficient large number of generations, which we set to be 3000.

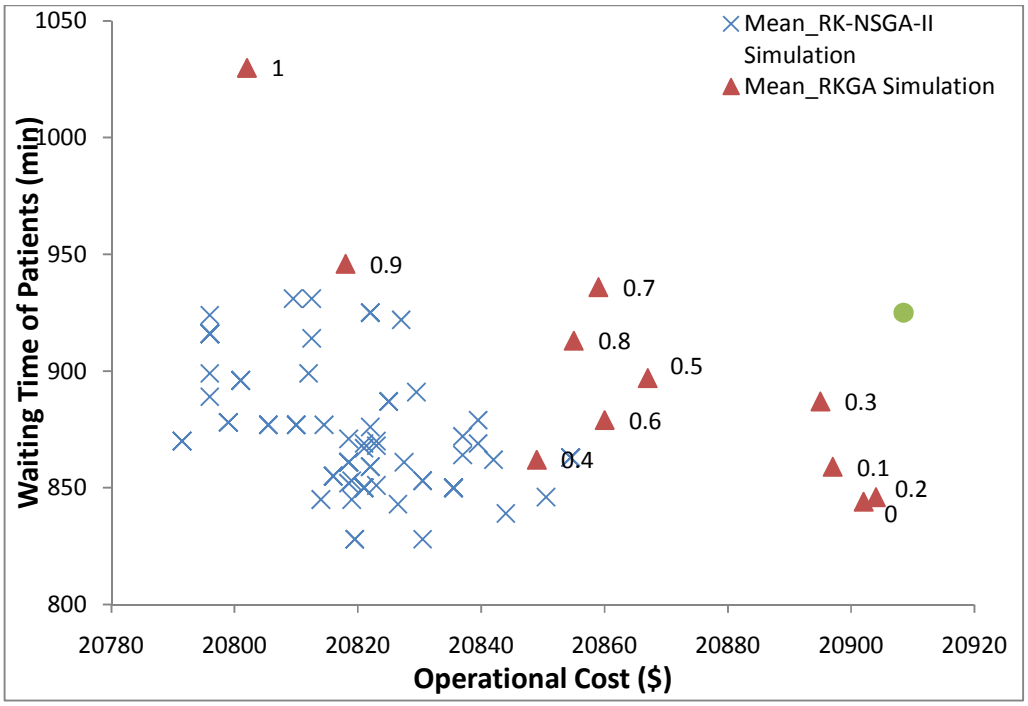
5.3.Testing the effectiveness by comparing with alternative approaches

The effectiveness of our approach is tested through comparison with single objective simulation optimization with GA operator and pure GA. In both

alternative methods, the random keys encoding structure is kept in the GA. In the RKGA simulation optimization, since RKGA cannot generate efficient frontier directly, we use relative weight factor of cost to from 0 to 1. In pure RKGA, only cost is used as the criteria as waiting time cannot be captured without simulation.

Population size, number of generations, crossover and mutation rates are all the same for all approaches. We use 75^{th} percentile of time distribution as the allocated duration for each specialty. After running all three approaches, the assignment and sequence of patients are obtained, which is input to a simulation model and compared performance using common random numbers. The performances of the approaches of three demand patterns can be found in Figure 14. There are two observations from the figure: (1) for two approaches both using simulation optimization, our approach using RK-NSGA-II as the optimizer is outperforming the RKGA optimizer, especially when the demand is larger; (2) for the two approaches both using RKGA as the optimization algorithm, RKGA simulation optimization is dominating the solution from pure RKGA, under all three demand patterns.


Min_RK-NSGA-II Simulation
Min_RKGA Simulation
Waiting Time of Patients (min)
Operational Cost ($)
1200
1100
1000
900
800
700
600
500
400
19000
20000
21000
22000
23000
24000
25000
1
0.9
0.8
0.7
0.6
0.4
0.5
0.3
0.1
0.2
0


Mean_RK-NSGA-II Simulation
Mean_RKGA Simulation
Waiting Time of Patients (min)
Operational Cost ($)
1050
1000
950
900
850
800
20780
20800
20820
20840
20860
20880
20900
20920
1
0.9
0.7
0.8
0.5
0.6
0.4
0.3
0.1
0.2
0

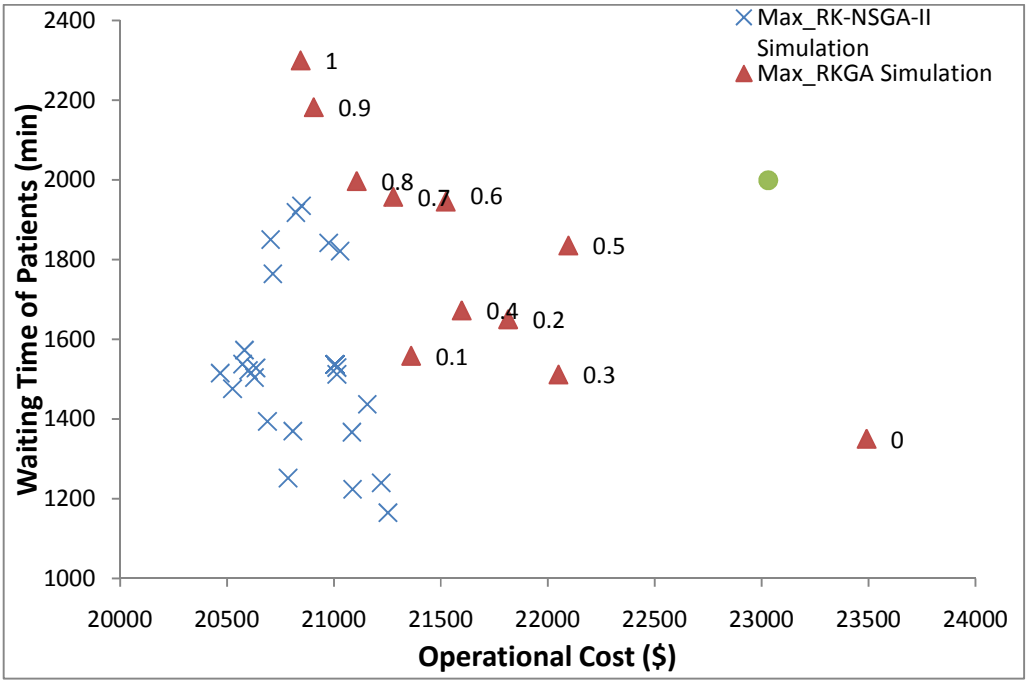


Figure 14. Comparison of three approaches

6. Investigating managerial questions

6.1. What is the optimal length of time block for each case?

We investigate the managerial questions mentioned in section 1, starting with testing different length of the time allocation for each surgery. Since the 65th percentile of case duration is very close to the mean duration in our sample data, in the experiment we include the 65th, 75th and 85th percentile of the case distribution. The basic structure is seen in all experiments with different demand patterns. The result from mean demand is shown in Figure 15. It can be seen that as the time allocation increases, the patient waiting time is decreasing, while the operational cost is increasing. There is no clear domination between different time allocations. Thus the decision maker can choose according to the hospital's own criteria.

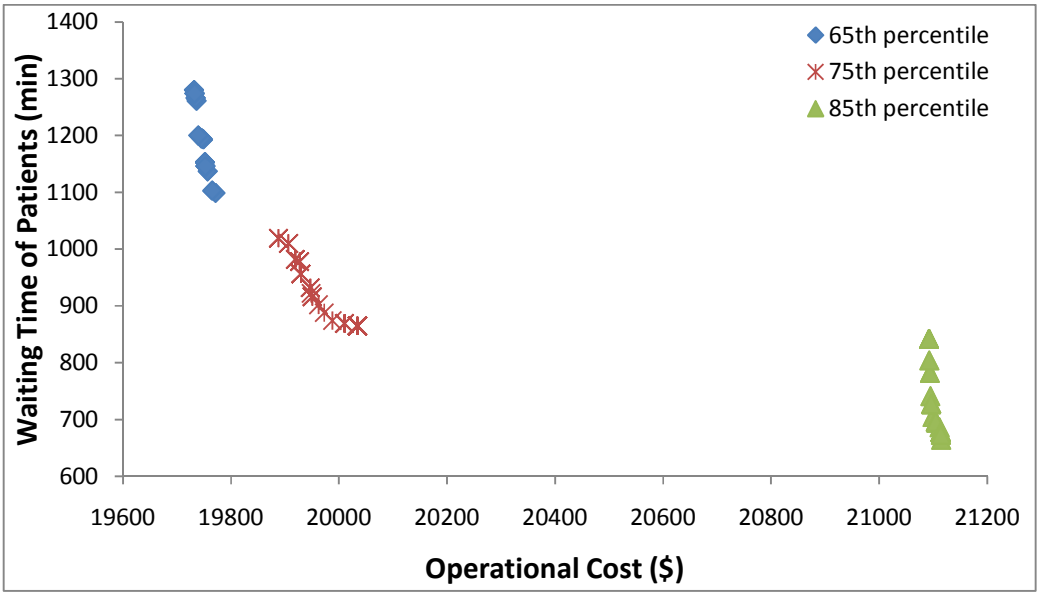


Figure 15. Pareto frontiers for allocating 65th, 75th & 85th percentile to surgeries

6.2.How much impact does patient no-show have on the scheduling performance?

We then use our approach to investigate the impact of no-show on the scheduling performance. In this design of experiment, no-show rate has five levels ranging from 0% to 20%; no-show occurrence has three types of distributions: all day, morning only, or afternoon only. The result is shown in Table 18 and Figure 16.

Table 18

Numerical results from the no-show impact analysis

No-show rate	*Time of no-show*	*Estimated average operational cost*	*Estimated average waiting time*
0%	N/A	3843.92	65.91
	All day	3870.39	57.53
5%	Morning	3895.47	51.60
	Afternoon	3847.90	62.90
	All day	3893.31	51.42
10%	Morning	3940.25	44.53
	Afternoon	3847.71	65.47
	All day	3918.22	49.83
15%	Morning	3990.84	37.19
	Afternoon	3866.03	61.49
	All day	3939.47	47.13
20%	Morning	4018.09	38.08
	Afternoon	3848.28	61.58

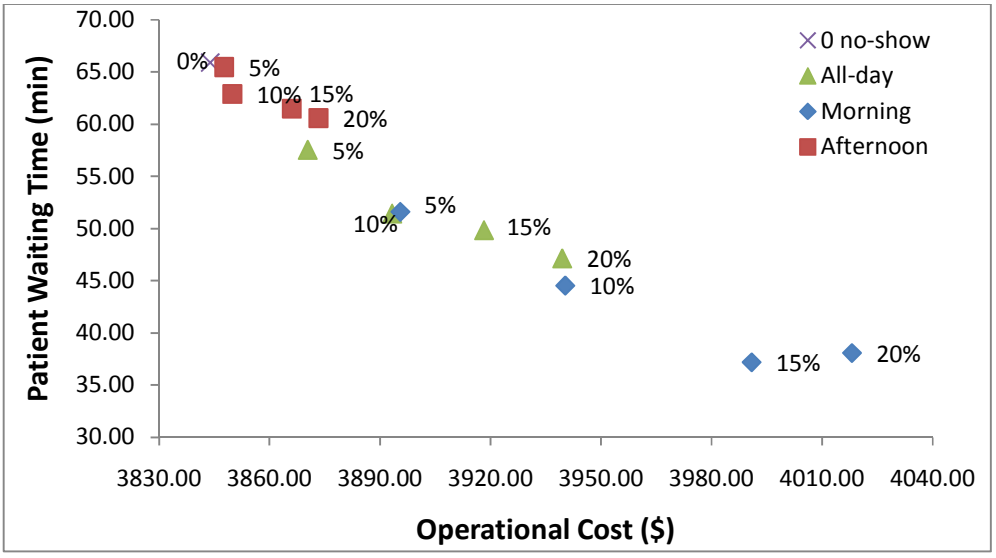


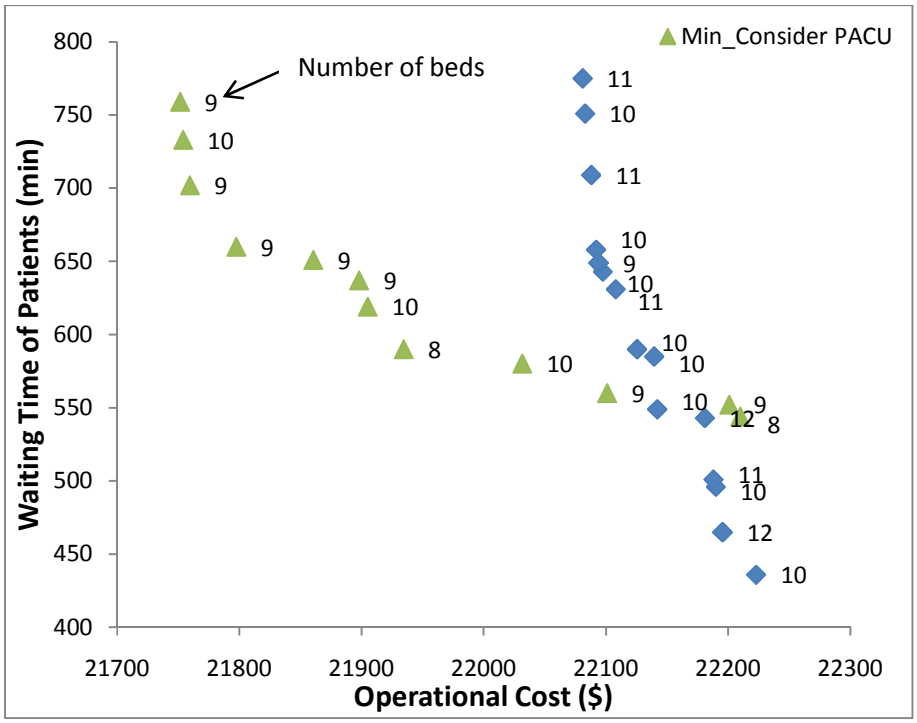
Figure 16. Results from no-show impact analysis

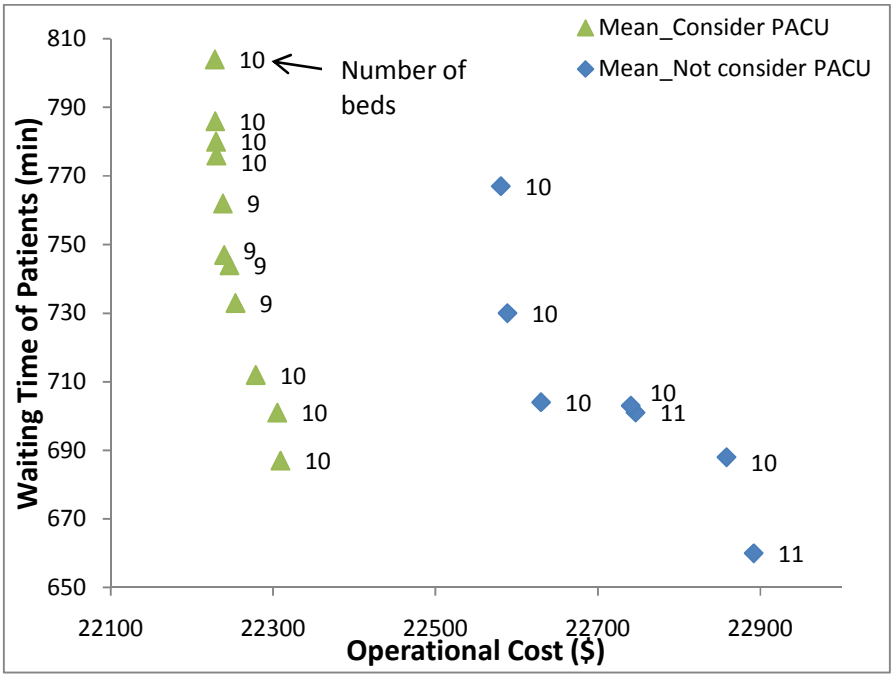
It is shown that as the no-show rate increases, cost is increasing and waiting time is decreasing. Within a day, the earlier the no-show is distributed, the higher the cost and the lower the waiting time. Since no-show rate may vary

among different specialties, surgical groups or patients (literature), the schedule can be adjusted accordingly in practice. Meanwhile, the no-show rate can be incorporated while deciding the time allocation of patients', which is also affecting both performance measures as indicated in section 6.1.

6.3.How much impact does the downstream resource have on the scheduling?

Next, our approach is used to find if there is statistical difference whether considering PACU simultaneously while scheduling or not. After conducting experiments for three different level of demand, the result is shown in Figure 17. The label of each point indicates the number of beds used in PACU for that solution. It is seen to be more costly not to consider PACU in all demand levels.





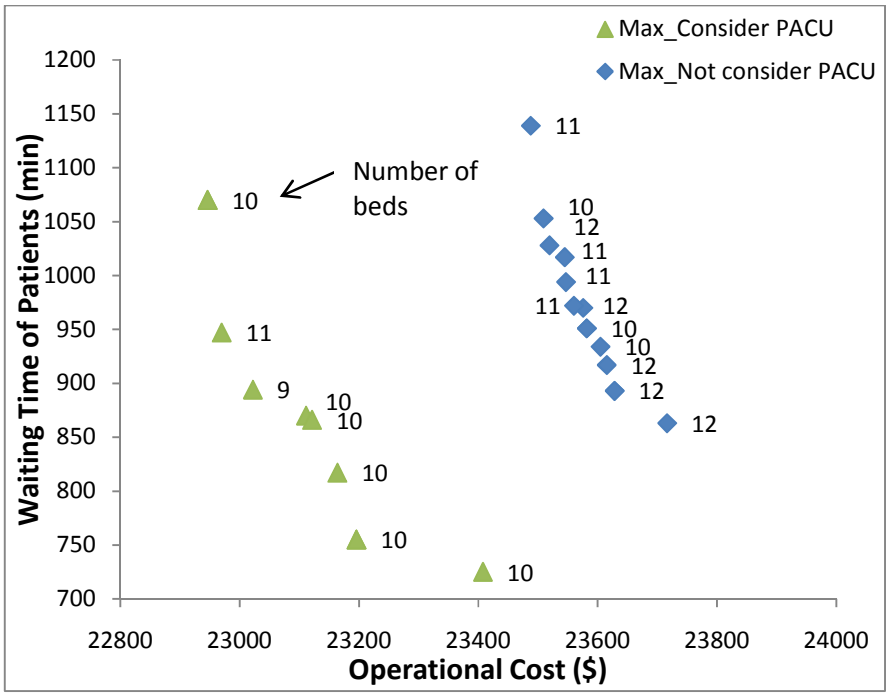


Figure 17. Comparison of the two scenarios

The estimated average of the number of PACU beds, cost and waiting time can be found in Table 19. The gap in performance measures between two scenarios is calculated. Although the number of beds has been decreased by more than 10% and shown to be statistically different with 99% confidence by considering PACU while scheduling, neither the cost nor the waiting time is affected statistically, either. Thus in practice, whether to consider PACU depends almost solely on the scarcity of PACU resource. If the number of beds or the staffing in PACU is limited, we suggest considering PACU while scheduling.

Table 19

Numerical results of the two scenarios

Demand	Criteria (Estimated average)	Consider PACU while scheduling	Not consider PACU while scheduling	Gap	Statistically different? (with 99% confidence)
Min	Number of beds	9.13	10.21	10.58%	Y
	Cost ($)	21918.4	22130.93	0.96%	N
	Waiting time (min)	638	607	-4.85%	N
Mean	Number of beds	9.51	10.85	12.35%	Y
	Cost ($)	22271.84	22767.62	2.18%	N
	Waiting time (min)	728	696	-4.50%	N
Max	Number of beds	9.92	11.5	13.74%	Y
	Cost ($)	23031.38	23582.7	2.34%	N
	Waiting time (min)	916	930	1.51 %	N

7. Conclusions

In this chapter, we develop a modeling framework based on simulation optimization to assist multi-criteria surgery scheduling according to hospital management desires. The Pareto frontier allows managers to make the best decisions. The integration of simulation and optimization incorporate more uncertainty factors than existing optimization methods. It was experimentally shown that our proposed RK-NSGA-II is an effective technique for finding Pareto optimal solutions which are found by 3000 generations.

Using our methodology, it is shown how the time allocation and no-show are affecting the scheduling performance. We also compare whether or not to consider PACU while scheduling. The results suggest that it is affecting the number of beds in PACU but not cost or patient waiting.

Future work could extend the model to explore more on the uncertainties in OR, such as the how to plan overbooking bearing the fact that a certain portion of patients will not show up.

CHAPTER 5

CONCLUDING REMARKS

In this dissertation we develop three models to assist the multi-objective decision making and analysis in OR scheduling using simulation, math programming, meta-heuristics and simulation optimization.

In chapter 2 we develop a simulation model to evaluate the efficiency of cath lab operations in a major local health care facility. We vary the key parameters in the model such as length of the time-block assigned to each case, length of lunch buffers as well as the option of rescheduling patients, and consider both operational costs and patient satisfaction metrics to illustrate the tradeoffs between the two. Detailed experimentation help recommend allocating to each case a time block equal to the 75^{th} percentile of the case duration distribution and scheduling a short buffer in the middle and at the end of each day to absorb variation and reduce the possibility of overtime. Sensitivity analysis is performed on key variables to test the robustness of our recommendations. Overall we find that the 75^{th} percentile of process duration is always on the efficient frontier and is a good compromise of both operational cost and patient waiting well. The health care facility adopted our recommendations and is now realizing the anticipated improvements.

Chapter 3 introduces a two-phase modeling approach to OR planning and scheduling. We considered multiple criteria while evaluating the performance of

the planning and scheduling, including OR utilization and overtime, surgeons' preference, patient waiting time and patient flow from OR to PACU. The exact solution from Phase 1 illustrates the trade-offs between operational objectives and patient/surgeon satisfaction objectives. Also it shows that fewer number of surgeon group yields better performances. In the Phase 2 optimal solution from MIP is compared with three heuristics and RKGA. By applying RKGA, close-to-optimal solutions can be derived much more efficiently. The impact of decomposing the problem is found to be within 11% to the optimality, which is tolerable considering the 99% saving in computation time.

Chapter 4 develops a modeling framework based on simulation optimization to schedule surgeries according to hospital management desires. We use RKGA and NSGA-II as the optimizer. The integration of simulation and optimization incorporate more uncertainty factors than existing optimization methods, and guides the simulation to optimum efficiently. The Pareto optimal set of solutions allows managers to trade-off between multi-criteria and make their best decisions. It is experimentally shown that our proposed RK-NSGA-II is an effective technique for finding Pareto optimal solutions which are found by 3000 generations.

Using our methodologies, hospital managers can allocate capacity and schedule patients with compromise to multiple objectives according to their own preference. In this dissertation we have also shown how to use the methodologies introduced to investigate managerial questions in the real world. For example, using the MIP formulation, we find that fewer groups with more surgeons in each

group is outperforming more groups with less surgeons; we compare whether or not to consider PACU while scheduling using the simulation optimization framework and results suggest that it is affecting the number of beds in PACU but not cost or patient waiting. By developing OR scheduling models, the managers are able to make multi-criteria decisions based on system-wide performance.

REFERENCES

Ahmed, M. A., & Alkhamis, T. M. (2009). Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research*, *198*, 936-942

Almeida, M. R. D., Pacheco, M. A. C., Hamacher, S., & Velasco, M. (2003). Optimizing the production scheduling of a petroleum refinery through genetic algorithms. *International Journal of Industrial Engineering: Theory, Applications and Practice*, *10*(1), 35-44

Allaoui, H. & Artiba, A. (2004). Integrating simulation and optimization to schedule a hybrid flow shop with maintenance constraints. *Computers & Industrial Engineering*, *47*, 431-450

Angelis, V. D., Felici, G., & Impelluso, P. (2003). Integrating simulation and optimization in health care center management. *European Journal of Operational Research*, *150*(1), 101-114

Baesler, F., & Sepulveda, J. (2000). Multi-response simulation optimization using stochastic genetic search within a goal programming framework. *Proceedings of the 2000 Winter Simulation Conference*, 788-794

Baesler, F., & Sepulveda, J. (2001). Multi-objective simulation optimization for a cancer treatment center. *Proceedings of the 2001 Winter Simulation Conference*, 1405-1411

Bean, J. C. (1984). Genetic algorithms and random keys for sequencing and optimization. *ORSA Journal on Computing*, 6, 154-160

Blake, J. T., & Carter, M. W. (2002). A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, *140*, 541-561

Blake, J. T., Dexter, F., & Donald, J. (2002). Operating room manager's use of integer programming for assigning block time to surgical groups: a case study. *Anesthesia and Analgesia*, *94*, 1272-1279

Boyle, C., & Shin, W. (1996). An interactive multiple-response simulation optimization method. *IIE Transactions*, *28*, 319-331

Cardoen, B., Demeulemeester, E., & Belien, J. (2010). Operating room planning and scheduling: a literature review. *European Journal of Operational Research*, *201*(3), 921-932

Cardoen, B., Demeulemeester, E., & Demeulemeester, E. (2007). Evaluating the capacity of clinical pathways through discrete-event simulation. Working Paper

Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and Operations management*, *12*, 519-549

Chernew, M. E., Hirth, R. A., & Cutler, D. M. (2009). Increased spending on health care: long-term implications for the nation. *Health Affairs*, *28*(5), 1253-1255

Davies, R., & Davies, H. (1994). Modelling patient flows and resource provision in health systems. *Omega*, *22*, 123-131

Davis, L. (1991). *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, New York, NY

De Angelis, V., Felici, G., & Impelluso, P. (2003). Integrating simulation and optimization in health care center management. *European Journal of Operational Research*, *105*, 101–114

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182-197

Denton, B., Miller, A., Balasubramania, H., & Huschka, T. (2009). Optimal allocation of surgery blocks operating rooms under uncertainty. Working paper

Denton, B., Viapiano, J., & Vogl, A. (2007). Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, *10*(13), 13-24

Dexter, F., Macario, A., Lubarsky, D. A., & Burns, D. D. (1999a). Statistical method to evaluate management strategies to decrease variability in operating room utilization: Application of linear statistical modeling and Monte-Carlo simulation to operating room management. *Anesthesiology*, *91*, 262-274

Dexter, F., Macario, A., Traub, R. D., Hopwood, M., & Lubarsky, D. A. (1999b). An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. *Anesthesia & Analgesia*, *89*, 7-20

Dexter, F., & Traub, R. D. (2002). How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia and analgesia*, *94*(4), 933-42

Erdogan, S.A., & Denton B. (2009). Surgery planning and scheduling: A literature review. Working paper

Eskandari, H., Rabelo, L., & Mollaghasemi, M. (2005). Multi-objective simulation optimization using an enhanced genetic algorithm. *Proceedings of the 2005 Winter Simulation Conference*, 833-841

Everett, J. E. (2002). A decision support simulation model for the management of an elective surgery waiting system. *Health care management science*, *5*(2), 89-95

Fei, H., Meskens, N., & Chu, C. (2006). An operating theatre planning and scheduling problem in the case of a "block scheduling" strategy. *Proceedings of the International Conference on Service Systems and Service Management*, 422-428

Fu, M. (2002). Optimization for simulation: theory vs. practice. *INFORMS Journal on Computing*, *14*(3), 192-215

Gerchak, Y., Gupta, D., & Mordechai, H. (1996). Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, *41*, 321-334

Glouberman S., & Mintzberg, H. (2001). Managing the care of health and the cure of disease--Part I: Differentiation. *Health Care Manage Review*, *26*, 56-69

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, & Machine Learning*. Addison-Wesley, Reading, MA.

Gupta, A.K. & Sivakumar, A.I. (2002). Simulation based multiobjective scheduling optimization in semiconductor manufacturing. *Proceedings of the 2002 Winter Simulation Conference*, 1862-1870

Gupta, D. (2007). Surgical Suites' Operations Management. *Society*, *16*(6), 689 - 700

Gupta, D., Natarajan, M. K., Gafni, A., Wang, L., Shilton, D., Holder, D., & Yusuf, S. (2007). Capacity planning for cardiac catheterization: a case study. *Health policy*, *82*, 1-11

Gupta, D., & Denton, B. (2008). Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, *40*(9), 800-819

Hans, E., Wullink, G., Vanhoudenhoven, M., & Kazemier, G. (2008). Robust surgery loading. *European Journal of Operational Research*, *185*(3), 1038-1050

Haral, U., Chen, R., Ferrell, W. G. Jr., & Kurz, M. B. (2006). Multiobjective single machine scheduling with nontraditional requirements. *International Journal of Production Economics*, *106*(2), 574-584

Holland, J. (1975). Adaptation in natural and artificial systems. *The University of Michigan Press*, 1975

Huang, H. C., Lee, L. H., Song, H., & Eck, B. T. (2009). SimMan - A simulation model for workforce capacity planning. *Computers & Operations Research*, *36*, 2490-2497

Huang, X. (1994). Patient Attitude Towards Waiting in an Outpatient Clinic and Its Applications. *Health Services Management Research*, *7*, 2-8

Jebali, A., Alouane, A. B. H., & Ladet, P. (2003). Performance comparison of two strategies for operating room scheduling. *Proceedings of the International Symposium on Computational Intelligence and Intelligent Informatics*.

Jebali, A., Hadjalouane, A., & Ladet, P. (2006). Operating rooms scheduling. *International Journal of Production Economics*, *99*(1-2), 52-62

Joines, J., Gokce, M., King, R., & Kay, M. (2002). Supply chain multi-objective simulation optimization. *Proceedings of the 2002 Winter Simulation Conference*, *1306-1314*

Jun, J. B., Jacobson, S. H., & Swisher, J. R. (1999). Application of discrete-event simulation in healthcare: a survey. *Journal of the Operational Research Society*, *50*, 109-123

Kaandorp, G. C., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science*, *10*, 217-229

Katzberg, R. W., & Haller C. (2006). Contrast-induced nephrotoxicity: Clinical landscape. *Kidney International*, *69*, S3-S7

Konak, A., Coit, D. W., & Smith, A. E. (2006). Multi-objective optimization using genetic algorithms: a tutorial. *Reliability Engineering and System Safety*, *91*, 992-1007

Krempels, K. H., Panchenko, A. (2006). An approach for automated surgery scheduling. *Proceedings of the Sixth International Conference on the Practice and Theory of Automated Timetabling*

Lee, L. H., Chew, E. P., Teng, S., & Goldsman, D. (2010). Finding the non-dominated Pareto set for multi-objective simulation models. *IIE Transactions*, *42*, 656-674

Lowery, J. C. (1998). Getting started in simulation in health care. *Proceedings of the 1998 Winter Simulation Conference*, *1*, 31-35

Lowery, J. C., & Davis, J. A. (1999). Determination of operating room requirements using simulation. *Proceedings of the 1999 Winter Simulation Conference*, 1568-1572

Macario, A., Vitez T. S., Dund B., & McDonale, T. (1995). Where are the costs in preoperative case? Analysis of hospital costs and charges for inpatient care. *Anesthesiology*, *83*(6), 1138-1144

Magerlein, J. M. & Martin, J. B. (1978). Surgical demand scheduling: A review. *Health Services Research*, *13*, 418-433

Marcon, E., & Dexter, F. (2006). Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Science*, *9*, 87-98

McIntosh, C., Dexter, F., & Epstein, R. H. (2006). The impact of service-specific staffing, case scheduling, turnovers, and first-case starts on anesthesia group and operating room productivity: a tutorial using data from an Australian hospital. *Anesthesia and analgesia*, *103*(6), 1499-1516

Michalewicz, Z. (2000). Evolutionary Computation 1: Basic Algorithms and Operators, 56-61.

Mollaghasemi, M., & Evans, G. (1994). Multicriteria design of manufacturing systems through simulation optimization. *IEEE Transactions on System, Man and Cybernetics*, *24*(9), 1407-1411

Mollaghasemi, M., Evans, G., & Biles, W. (1991). An approach for optimizing multi-response simulation models. *Computers & Industrial Engineering*, *21*, 201-203

Moving up the agenda. (2009, June 6th-12th). *The Economist.*

Mullen, P. M. (2003). Prioritizing waiting lists: how and why? *European Journal of Operational Research*, *150*, 32-45

Ogulata, S. N. & Erol, R. (2003). A hierarchical multiple criteria mathematical programming approach for scheduling general surgery operations in large hospitals. *Journal of Medical Systems*, *27*(3), 259-270

Ozcan, Y. A. (2005). Quantitative methods in health care management.

Persson, M., & Persson, J. A. (2009). Health economic modeling to support surgery management at a Swedish hospital. *Omega*, *37*, 853-863

Pham, D.-N. & Klinkert, A. (2008). Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, *185*(3), 1011-1025

Rohleder, T. R., & Klassen, K. J. (2002) Rolling horizon appointment scheduling: A simulation study. *Health Care Management Science*, *5*, 201-209

Roland, B., Martinelly, C., & Riane, F. (2006). Operating Theatre Optimization : A Resource-Constrained Based Solving Approach. *2006 International Conference on Service Systems and Service Management*, 443-448

Romanin-Jacur, G., & Facchin, P. (1987). Optimal planning of a pediatric semi-intensive care unit via simulation. *European Journal of Operational Research*, *29*, 192-198

Sandberg, W. S., Daily, B., Egan, M., Stahl, J. E., Goldman, J., Wiklund, R., Rattner, & R. W. (2005). Deliberate perioperative systems design improves operating room throughput. *Anesthesiology*, *103*(2), 406-408

Sciomachen, A., Tanfani, E., & Testi, A. (2005). Simulation models for optimal schedules of operating theatres. *International Journal of Simulation*, *6*(12-13), 26-34

Spencer, F. A., Goldberg, R. J., Becker, R. C., & Gore, J. M. (1998). Seasonal distribution of acute myocardial infarction in the second National Registry of Myocardial Infarction. *Journal of American College of Cardiology*, *31*, 1226-1233

Strum, D. P., Vargas, L. G., & May, J. H. (1999). Surgical subspecialty block utilization and capacity planning: A minimal cost analysis model. *Anesthesiology*, *90*, 1176-1185

Swisher, J. R., Jacobson, S. H., Jun, B., & Balci, O. (2001). Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & Operations research*, *28*, 105-125

Teleb, R., & Azadivar, F. (1994). A methodology for solving multi-objective simulation-optimization problems. *European Journal of Operational Research*, *72*, 132-145

Testi, A., Tanfani, E., Torre, G. (2007). A three-phase approach for operating theatre schedules. *Health Care Management Science*, *10*, 163-172

VanBerkel, P. T., & Blake, J. T. (2007). A comprehensive simulation for wait time reduction and capacity planning applied in general surgery. *Health Care Management Science*, *10*, 373-385

Vanden Bosch, P. M., & Dietz, D. C. (2000). Minimizing expected waiting in a medical appointment system. *IIE Transactions*, *32*(9), 841-848

Willis, K. O., & Jones, D. F. (2008). Multi-objective simulation optimization through search heuristics and rational database analysis. *Decision Support Systems*, *46*, 277-286

Zakerifar, M., Biles, W. E., & Evans, G. W. (2009). Kriging metamodeling in multi-objective simulation optimization. *Proceedings of the 2009 Winter Simulation Conference*, 2115-2122

APPENDIX A

COMBINED MODEL OF PHASE 1.2 AND PHASE 2

Notations and decision variables:

$t \in T$	index of day
$i \in I$	index of room block ($I = 2N \times T$)
λ_1	relative weight factor of cost
c_b	cost of staffing in PACU
Z_{rpi}	1 if patient *p* is assigned to surgeon *r* in room block *i*, 0 otherwise
U_i^-	under-utilization in room block *i*
U_i^+	over-utilization in room block *i*
W_p	waiting time of patient *p* on the waiting list after the due date
BM_t	number of beds in PACU on day *t*
$X1_{pi}$	start time of patient *p* in OR in room block *i*
$X2_{pt}$	start time of patient *p* in PACU on day *t*
$X3_{pt}$	start time of patient *p* in OR on day *t*
$OR_{pp'}$	1 if patient *p* proceeds patient *p'* in the same OR, 0 otherwise
$PACU_{pp'}$	1 if patient *p* proceeds patient *p'* in the same PACU bed, 0 otherwise
F_{bpt}	1 if patient *p* is assigned to bed *b* on day *t*, 0 otherwise

MIP formulation of combined two phases:

$$\textbf{Min}\ \lambda_1 \cdot \left(\sum_i \left(c_1 \cdot U_i^+ + c_2 \cdot U_i^- \right) + \sum_t c_b \cdot BM_t \right) + \left(1-\lambda_1\right) \cdot \sum_p W_p \tag{1}$$

$$Z_{rpi} \le \sum_s \left(pts_{ps} \cdot \sum_m \left(y_{mi} \cdot surs_{ms} \cdot surg_{rm} \cdot sa_{ri} \right) \right) \qquad \forall r \in R, p \in P, t \in T \tag{2}$$

$$\sum_i Z_{rpi} \le pg_{pr} \qquad \forall r \in R, p \in P \tag{3}$$

$$\sum_r \sum_i Z_{rpi} \le 1 \qquad \forall p \in P \tag{4}$$

$$\sum_m \sum_p \sum_r \left(Z_{rpi} \cdot y_{mi} \cdot surg_{rm} \cdot \tau_p \right) + e_i \le cap_i + brk_i + \frac{ot}{2} \qquad \forall i \in I \tag{5}$$

$$\sum_m \sum_p \sum_r \left(\left(Z_{rpi} \cdot y_{mi} \cdot surg_{rm} \cdot \tau_p \right) + \left(Z_{rp(i+1)} \cdot y_{m(i+1)} \cdot surg_{rm} \cdot \tau_p \right) \right) + e_i + e_{i+1}$$

$$\le cap_i + cap_{i+1} + brk_i + brk_{i+1} + ot \qquad \forall i \in \{1,3,5,\ldots I-1\}, p \in P, t \in T \tag{6}$$

$$ut \cdot ra_i - \sum_m \sum_p \sum_r \left(Z_{rpi} \cdot y_{mi} \cdot surg_{rm} \cdot \tau_p \right) - e_i \le U_i^- \qquad \forall i \in I, t \in T \tag{7}$$

$$\sum_m \sum_p \sum_r \left(Z_{rpi} \cdot y_{mi} \cdot surg_{rm} \cdot \tau_p \right) + e_i - cap_i - brk_i \le U_i^+ \qquad \forall i \in I, t \in T \tag{8}$$

$$\sum_r \sum_i \left\lceil \frac{i}{2N} \right\rceil \cdot Z_{rpi} \le W_p \qquad \forall p \in P \tag{9}$$

$$\left(\frac{T}{2} + 1 - DD_p \right) \cdot \left(1 - \sum_r \sum_i Z_{rpi} \right) \le W_p \qquad \forall p \in P \tag{10}$$

$$\sum_r r \cdot Z_{rpi} \le \sum_r r \cdot Z_{rp'i} + l \cdot \left(1 - \sum_r Z_{rp'i} \right) \qquad \forall i \in I, p \in P, p' \in P, p \ne p' \tag{11}$$

$$X1_{pi} \le l \cdot \sum_r Z_{rpi} \qquad \forall i \in I, p \in P \tag{12}$$

$$X2_{pt} \le l \cdot \sum_r \sum_{i=2Nt}^{2N(t+1)} l \cdot Z_{rpi} \qquad \forall p \in P, t \in T \tag{13}$$

$$X3_{pt} + l \cdot \sum_{r} l \cdot Z_{rpi} \le X1_{pi} \qquad \forall i \in \{2Nt,...,2N(t+1)\}, p \in P, t \in T \tag{14}$$

$$X2_{pt} = X3_{pt} + \tau_p \cdot \sum_{r} \sum_{i=2Nt}^{2N(t+1)} Z_{rpi} \qquad \forall p \in P, t \in T \tag{15}$$

$$X1_{pi} + \tau_p \le X1_{p'i} + l \cdot \left(3 - OR_{pp'} - \sum_{r} Z_{rpi} - \sum_{r} Z_{rp'i} \right)$$

$$\forall i \in I, p \in P, p' \in P, p \ne p' \tag{16}$$

$$X1_{p'i} + \tau_{p'} \le X1_{pi} + l \cdot \left(2 + OR_{pp'} - \sum_{r} Z_{rpi} - \sum_{r} Z_{rp'i} \right)$$

$$\forall i \in I, p \in P, p' \in P, p \ne p' \tag{17}$$

$$X2_{pt} + \upsilon_p \le X2_{p't} + l \cdot \left(3 - PACU_{pp'} - F_{bpt} - F_{bp't} \right)$$

$$\forall b \in B, t \in T, p \in P, p' \in P, p \ne p' \tag{18}$$

$$X2_{p't} + \upsilon_{p'} \le X2_{pt} + l \cdot \left(2 + PACU_{pp'} - F_{bpt} - F_{bp't} \right)$$

$$\forall b \in B, t \in T, p \in P, p' \in P, p \ne p' \tag{19}$$

$$b \cdot F_{bpt} \le BM_t \qquad \forall b \in B, p \in P, t \in T \tag{20}$$

$$\sum_{r} \sum_{i=2Nt}^{2N(t+1)} l \cdot Z_{rpi} = \sum_{b} F_{bpt} \qquad \forall p \in P, t \in T \tag{21}$$

$$X1_{pi}, X2_{pt}, X3_{pt}, W_p, BM_t, U_i^+, U_i^- \ge 0 \qquad \forall p \in P, i \in I, t \in T \tag{22}$$

$$Z_{rpi}, F_{bp}, OR_{pp'}, PACU_{pp'} \in \{0,1\} \qquad \forall b \in B, r \in R, i \in I, p \in P, p' \in P \tag{23}$$

The objective (1) minimizes the weighted total of cost and patients' waiting time. The cost is composed of the cost of under- and over-utilization in the OR, and the maximum number of beds in PACU on each day. Constraints (2) to (11) are the constraints from Phase 1.2. (12) to (22) are from Phase 2.

Specifically, constraints (2) ensure that each patient is assigned to at most one surgeon in the same specialty, only if the surgical group that the surgeon is in is assigned with that block. Constraints (3) indicate that each patient must be assigned to a surgeon that can be assigned to this patient. Constraints (4) guarantee that each patient is assigned to at most one surgeon at a time. Constraints (5) and (6) ensure that the operating room is scheduled within the capacity and overtime limit in each individual block and each day, respectively. Constraints (7) define the under-utilization as the difference between the target usage and the total surgery time in a block, if the operating room is available and there is under-utilization. Constraints (8) define the over-utilization as the difference between the sum of all surgery time and the sum of capacity and lunch break of a block, if there is overtime. For a *scheduled* patient *p*, if scheduled after due date, the waiting time on the waiting list is defined in constraints (9) as the difference between the date that he/she is scheduled and the due date. For an *unscheduled* patient *p*, since he/she will be scheduled at least one day after the planning horizon $T/2$, the waiting time is defined in constraints (10) as the difference between $(T/2+1)$ and the due date. Constraints (11) guarantee that one surgeon can work in at most one room at a time. Constraints (12), (13) and (14) ensure that the starting time of a patient in OR and PACU on a day to be zero if the patient is not scheduled on that day. Constraints (15) guarantee that the starting time of a patient in PACU is the estimated surgery time of that patient in addition to his/her starting time in OR. Constraints (16) and (17) ensure that OR cannot be occupied by more than one patient at a time. Constraints (18) and (19)

ensure that a bed in PACU cannot be occupied by more than one patient at a time. Constraints (20) indicate that the assigned index of beds on each day have to be less than the maximum number of beds in PACU of that day. Constraints (21) indicate that all patients should have the surgery and recovery on the same day.