Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

PERFORMANCE ANALYSIS AND SCHEDULING STRATEGIES FOR AMBULATORY SURGICAL FACILITIES

by
Xuanqi Zhang

Ambulatory surgery is a procedure that does not require an overnight hospital stay and is cost effective and efficient. The goal of this research is to develop an ASF operational model which allows management to make key decisions. This research develops and utilizes the simulation software ARENA based model to accommodate: (a) Time related uncertainties – Three system uncertainties characterize the problem (ii) Surgery time variance (ii) Physician arrival delay and (iii) Patient arrival delay; (b) Resource Capture Complexities – Patient flows vary significantly and capture/utilize both staffing and/or physical resources at different points and varying levels; and (c) Processing Time Differences – Patient care activities and surgical operation times vary by type and have a high level of variance between patient acuity within the same surgery type. A multi-dimensional ASF non-clinical performance objective is formulated and includes: (i) Fixed Labor Costs – regular time staffing costs for two nurse groups and medical/tech assistants, (ii) Overtime Labor Costs – staffing costs beyond the regular schedule, (iii) Patient Delay Penalty – Imputed costs of waiting time experienced patients, and (iv) Physician Delay Penalty – Imputed costs of physicians having to delay surgical procedures due to ASF causes (limited staffing, patient delays, blocked OR, etc.).

Three ASF decision problems are studied: (i) Optimize Staffing Resources Levels - Variations in staffing levels though are inversely related to patient waiting times and physician delays. The decision variable is the number of staff for three resource groups,

for a given physician assignment and surgery profile. The results show that the decision space is convex, but decision robustness varies by problem type. For the problems studied the optimal levels provided 9% to 28% improvements relative to the baseline staffing level. The convergence rate is highest for less than optimal levels of Nurse-A. The problem is thus amenable to a gradient based search. (ii) Physician Block Assignment - The decision variables are the block assignments and the patient arrivals by type in each block. Five block assignment heuristics are developed and evaluated. Heuristic #4 which utilizes robust activity estimates (75% likelihood) and generates an asymmetrical resource utilization schedule, is found to be statistically better or equivalent to all other heuristics for 9 out of the 10 problems and (iii) Patient Arrival Schedule – Three decision variables in the patient arrival control (a) Arrival time of first patient in a block (b) The distribution and sequence of patients for each surgery type within the assigned windows and (c) The inter arrival time between patients, which could be constant or varying. Seven scheduling heuristics were developed and tested. Two heuristics one based on Palmers Rule and the other based on the SPT (Shortest Processing Time) Rule gave very strong results.

PERFORMANCE ANALYSIS AND SCHEDULING STRATEGIES FOR AMBULATORY SURGICAL FACILITIES

by
Xuanqi Zhang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Industrial Engineering

Department of Mechanical and Industrial Engineering

January 2014

.

APPROVAL PAGE

PERFORMANCE ANALYSIS AND SCHEDULING STRATEGIES FOR AMBULATORY SURGICAL FACILITIES

Xuanqi Zhang

Dr. Sanchoy K. Das, Dissertation Advisor — Date
Professor & Graduate Advisor IE, MNE, HSM & PS, NJIT

Dr. Reggie J. Caudill, Committee Member — Date
Professor and Chair of Mechanical and Industrial Engineering, NJIT

Dr. Golgen Bengu, Committee Member — Date
Associate Professor of Mechanical and Industrial Engineering, NJIT

Dr. Wenbo Cai, Committee Member — Date
Assistant Professor of Mechanical Engineering, NJIT

Dr. Cheickna Sylla, Committee Member — Date
Professor of Decision Sciences & MIS, NJIT

BIOGRAPHICAL SKETCH

Author: Xuanqi Zhang

Degree: Doctor of Philosophy

Date: January 2014

Undergraduate and Graduate Education:

- Doctor of Philosophy in Industrial Engineering,
  New Jersey Institute of Technology, Newark, NJ, 2014

- Bachelor of Science in Industrial Engineering,
  Huazhong University of Sci and Tech, Wuhan, P. R. China, 2010

Major: Industrial Engineering

Presentations and Publications:

Zhang, X. and Das, S. (2012), "Scheduling Capacity to Physician Groups in Ambulatory Surgical Facilities" Institute of Industrial Engineers (IIE) Annual Conference, May 19-23, Orlando, Florida. – Journal paper in process for IIE Transactions Healthcare Systems.

Zhang, X. and Das, S. (2013), "Deriving Optimal Nurse Staffing Level for a Ambulatory Surgical Facility – A Simulation Approach" Journal paper in process for IIE Transactions Healthcare Systems.

Das, S., Desai, M. and Zhang, X. (2010), "Healthcare Process Mapping: An Orthopedics Case Study", POMS Conference, College of Healthcare Operations Management (CHOM), Vancouver, Canada

Zhang, X. (2013), "A Simulation Analysis of Physician Surgery Start Delays in Ambulatory Surgical Centers." Dana Knox Student Research Showcase, New Jersey Institute of Technology.

I dedicate my dissertation work to my family and many friends. There is a special feeling of gratitude to my loving parents: my father, Lanlei Zhang（张兰雷） and my mother, Yuxia Deng (邓宇霞), whose words of encouragement and caring have made me never give up. Some special thanks also to my grandparents: Zizhong Zhang (张自忠), Guofeng Lan （兰国凤）, Benqin Deng （邓本勤） and Kerong Rao （饶克荣） who taught me many principles of life when I was a child. And also thanks to my relatives' concerns for me and my big family.

I also dedicate this dissertation to my dear friends and church family who have supported me throughout the process. I will always appreciate all they have done, especially John (Jack) Gidney for helping me with my Teaching Assistant experiments demonstrations, my colleagues: Dr. Shivon Boodhoo, Zhenqing Zheng, and Yifeng Liu for the many hours of research advising and discussion, and Reagan Randall, Ghafor Vala, Zimeng Cheng and Yixuan Li for helping me adapt into foreign life and study quickly at my first year in the US.

I dedicate this work and give special thanks to my best friends: Mengqi Yuan, Qian Li and my wonderful college roommates: Lianglu Che, Rui Zhou and Ting Zheng for being there for me throughout the entire doctorate program though far away. All my classmates in class 0013, class 0329 and IE 0602, my dear friends, current colleagues and people I met in my life, all of you have given me lessons to bring me here.

Thanks to all of you being part of my life! And thanks to God!

ACKNOWLEDGMENT

It would not have been possible to write this dissertation without the help of my dear advisor, Professor Sanchoy Das, who helped me in the research path from beginning to now with his caring, patient, inspiring and knowledgeable guidance. He gave me opportunities to improve myself in the PhD study. Lessons taken from him cannot only bring me to who I am now but also lead me to my future life.

I would also like to thank my dissertation committee members: Professor Reggie Caudill, who was also my teacher in both simulation and cost management class, for giving me great advice on both courses and dissertation; Professor Golgen Bengu, who taught me statistics where I learned analysis methods which have been applied in my dissertation experiment results analysis report; Professor Wenbo Cai, for giving me encouragement when I was looking for effective heuristic algorithms for my dissertation; and Professor Cheickna Sylla, for being my committee member and giving me valuable comments from a different perspective.

My sincere thanks to Help Desk staff, Athelstan Nelson who helped me with software installation, update and system compatible issues, and to my colleagues: Dr. Shivon Boodhoo and Zhenqing Zheng who helped me with dissertation writing and presentation.

I would also thank my Chinese teacher Xiqiang Wang in high school for cheering me on for my college entrance exam, and my college teachers: Professor Liang Gao, Professor Zailin Guan, and Professor Qiong Liu for enlightening me in the first glance of research.

Finally, I would like to thank my parents: my father, Lanlei Zhang and my mother, Yuxia Deng, they were always there being supportive and stood by me through the good times and bad.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Research Background

Ambulatory or outpatient surgery is a surgical procedure that does not require an overnight hospital stay. Ambulatory surgery is a cost effective for providers and provides patients with effective and efficient care, making it one of the fastest growing segments of the US Healthcare system. Ambulatory surgical facilities (ASFs) provide the surgical facilities and associated staffing resources, while physician groups who contract with the ASF provide the patients and perform specific surgeries. ASFs are thus challenged to keep operating costs down, while at the same time keeping both physician and patient groups satisfied. Some key facts about ambulatory facilities are (all the following information is from Ambulatory Surgery in the United States, 2006, NHS) in:

- In 2006, an estimated 53.3 million procedures were performed during 34.7 million ambulatory surgery visits to 7000 different facilities in the US.

- Average times for surgical visits were higher for hospital-based centers than for visits to freestanding ambulatory surgery centers (147 minutes compared with 98 minutes).

- Frequently performed procedures included endoscopy of large intestine (5.7 million), endoscopy of the small intestine (3.5 million), extraction of lens (3.1 million), injection of agent into spinal canal (2.0 million), and insertion of prosthetic lens (2.6 million)

- ASFs allow surgeons to perform cases more efficiently. One study comparing spine procedures performed at hospitals and ASFs found 20% less time spent in the operating room. The turnaround time between procedures is also significantly less at an ASF than at a hospital. One spine surgeon found that the turnaround time between procedures at his ASF is 12 minutes, compared to a turnaround time of 1 hour and twenty 20 minutes at the local hospital.

- ASFs are a key part of the national healthcare cost reduction initiative. Currently, Medicare pays ASFs 58% of the amount paid to hospital outpatient for performing the same services. For example, Medicare pays hospitals $1,670 for performing an outpatient cataract surgery while paying ASFs only $964 for performing the same surgery.

ASFs are structurally complex in that patient volumes are dependent on independent physician groups. The ASFs do not directly recruit patients. So the ASF must satisfy both patient and physician groups while at the same time reducing labor cost, their primary cost driver. Currently the performance relationship between these three groups, and the sensibility to schedule and other operating parameter is unknown. ASFs thus use a variety of experience based trial-and-error to improve performance. A simulation is an effective approach to characterize system dynamic, and create algorithm solution for schedule, control staffing level, real time adjustment, and other ASF domains. Especially there is a need for improving the current resource (including staffing and facility) utilization and physician groups' schedule flexibility to get higher performance with lower cost. More accurate appointment schedules and more uncertainties within patients' appointment in ASFs are in need as well. Figure 1.1 lists the most common procedures performed in ASFs.

ASFs also require significant capital investment. Usually running about $1 million per OR, a small, single-specialty center with two surgical suites ranges from $2 million to $3 million, with larger-multispecialty ASFs costing $4 million to $8 million, according to calculations provided by Meridian Surgical Partners, which partners with physicians seeking to develop new ASFs in addition to acquiring interests in existing physician-owned facilities. As a result any costs saving can have a significant impact on ASF operations.

Surgical service	2007 Percent of volume	2007 Rank	2010 Percent of volume	2010 Rank
Cataract surgery w/ IOL insert, 1 stage	19.9%	1	17.6%	1
Upper GI endoscopy, biopsy	7.9	2	8.0	2
Diagnostic colonoscopy	5.9	3	4.2	5
Colonoscopy and biopsy	5.5	4	5.6	3
After cataract laser surgery	5.4	5	4.0	6
Lesion removal colonoscopy, snare technique	4.8	6	4.3	4
Injection spine: lumbar, sacral (caudal)	4.3	7	3.5	8
Injection foramen epidural: lumbar, sacral	3.1	8	3.8	7
Injection paravertebral: lumbar, sacral add on*	2.9	9	1.9	11
Injection paravertebral: lumbar, sacral*	1.9	10	2.1	9
Lesion removal colonoscopy, by biopsy forceps or bipolar cautery	1.7	11	1.1	17
Colon cancer screen, not high-risk individual	1.7	12	1.3	15
Injection foramen epidural add on	1.6	13	2.0	10
Upper GI endoscopy, diagnosis	1.5	14	1.3	16
Colorectal screen, high-risk individual	1.4	15	1.7	12
Cystoscopy	1.3	16	1.1	19
Destruction paravertebral nerve, add on	1.1	17	1.5	13
Revision of upper eyelid	0.9	18	1.0	20
Cataract surgery, complex	0.9	19	1.3	14
Injection spine: cervical or thoracic	0.8	20	0.8	26
Total	74.6		68.0	

Figure 1.1 Volume of surgical procedures in ASFs.

Source: National Health Statistics Reports Number 11 January 28, 2009–Revised

Healthcare systems typically involve multiple patient flow pathways, and tend not to be amenable to exact modeling methods. The literature demonstrates that simulation modeling is an effective and popular approach in healthcare analysis. This research also uses an ARENA based simulation model of a ASF as the primary analytical platform. The specific objectives of this research are: (i) Characterize and build a simulation model to represent the operating behavior of ambulatory surgical facilities (ASF), and use it to study performance sensitivity to key parameters such as capacity loading, physician assignment, staffing levels and patient arrival schedules; (ii) Develop a simulation experimental search procedure to derive the optimal ASF staffing strategy (nursing levels

and medical assistant staffing levels) for a given daily schedule and physician assignment; (iii) Develop a heuristic procedure(s) for generating the physician assignment including the specification of schedule blocks, surgery type balancing, and patient arrival rates. ASF would use this procedure (medium term intervals) to negotiate with physician groups. Objective is to optimize ASF performance as estimated by the simulation model; (iv) Develop a heuristic procedure to generate the daily patient arrival schedule based on surgery profile for the specific day. Objective is to minimize patient waiting time, without effective ASF performance. ASFs frequently experience extreme events, which are the common cause of performance slack. This dissertation will report in detail on the work associated with all objectives above.

1.2 Research Objectives and Accomplishments

This research is organized into the four research objectives described below. For each objective the accomplishments described in the subsequent chapters is briefly summarized.

1. Characterize and build a simulation model to represent the operating behavior of ambulatory surgical facilities (ASF), and use it to study performance behavior as a function of key parameters such as (i) capacity loading (ii) physician assignment (iii) staffing levels and (iv) patient arrival schedules.

Accomplishments: Field research of current operational flows of ASFs was done to build typical operations process diagram. Activities included (i) Direct Work Study

(ii) Discussions/Interviews with ASF Staff and (iii) Review of ASF Operations as Reported in the Literature. A generalized ASF process flowchart which identified (i) Patient transfer logic (ii) Resource usage profiles and (iii) Physician schedules and patient relationships has been created. A novel ASF operations objective function (non-clinical) which models (i) Regular and overtime staffing costs (ii) Patient waiting time costs and (iii) Physician delay costs has been formulated. Built and validated the corresponding ASF simulation model in the ARENA platform. Model was populated with reliable surgery and associated times (mean and standard deviations) allowing for accurate estimates that capture all systems variances.

2. Define and optimize the ASF staffing resource problem. Staffing costs are the largest direct cost of an ASF and the primary operations objective of ASF managers. Current practice, involves manual expertise whereby a person with staffing experience will make decisions on staff levels for the upcoming week. ASF operators need decision models that can characterize the relationship between staffing levels and operating costs, and consequently prescribe optimal staffing levels.

Accomplishments: Introduced the decision space as the staffing level for Nurse-A, Nurse-B and Medical Assistant, which are inversely related to two objective function terms: patient waiting times and physician delays. Since an analytical technique is inapplicable, a simulation based optimization approach was used to solve the problem. Two-dimensional convexity (Nurse-A and Nurse-B levels) of the objective

function is demonstrated for several test problems, confirming that a gradient search method can be efficiently used. The convergence rate is consistently highest for Nurse-A at the lowest staffing levels. Also shown is that the robustness of the decision space is not consistent across the problems. Variance analysis indicates a large performance range due the systemic combinations of the multiple variance sources in the ASF. Indicating that even with an optimal policy, major differences could be seen from day to day.

3. Defining and solve the physician block assignment problem. ASFs have the flexibility to decide how to assign schedule blocks (3 to 4 hour windows) to the different physician groups. The assignment solution affects the overall performance of the ASF for a multiple reasons including the combinatorial effect of the surgery types, surgery time variances, and resources requirements. ASFs need the assignment problem to be formalized and readily applicable solutions to be developed.

Accomplishments: Formulated the physician block assignment problem for the fixed staffing level case, as having two decision variables (i) Physician group is assigned to one or more continuous schedule block sufficient to meet their capacity needs and (ii) Number of patients for each surgery type scheduled to arrive in a block. Combining classical machine scheduling and assembly line balancing methods, several solution heuristics were developed. A theory of constraints approach was used to estimate robust process times. Heuristics were evaluated on a benchmark set of 10 problems

using the simulation model. Three heuristics are statistically dominant across the set of benchmark problems, with one proving the best solution for 9 of the 10 problems. The asymmetrical load balancing strategy is shown to be clearly effective in improving the ASF operation performance. A realistic lower bound was also derived, and for two problems the performance gap was about 20% indicating room for further improvement.

4. Defining and solving the ASF patient arrival time scheduling problem. Due to the order of magnitude difference between physician delay costs and patient delay costs, healthcare facilities in general schedule all patients to arrive much earlier than needed. There is now much research interest in developing patient arrival scheduling models. Specifically, ASFs needed models which consider patient surgery types and the associated physician group in generating an arrival schedule.

Accomplishments: Formulated the patient arrival time scheduling problem for the fixed staffing level case, as having three decision sets (i) identifying the patient arrival sequences for each group (ii) dynamic setting of the inter arrival time between every pair of patients and (ii) prescribing the arrival time of the first patient in each time block. Several solution heuristics were developed utilizing classical 3-machine sequencing methods such as the Cambell-Dudek-Smith and Palmer heuristics. Heuristics were evaluated on a benchmark set of 10 problems using the simulation model.

Figure 1.2 below shows the interrelationships between four the objectives. Clearly, the simulation model is central to this research since it provides the key estimates of physician delay, patient delay and staffing overtime. The three analytical models each make a decision utilizing a heuristic approach, but the quality of these decisions can only be assessed from the simulation model. In objectives #2 to #4 the model decisions for each experiment are entered into the simulation model which then estimates the performance variables. This in turn provides the objective function value which is the key to the evaluation analysis.

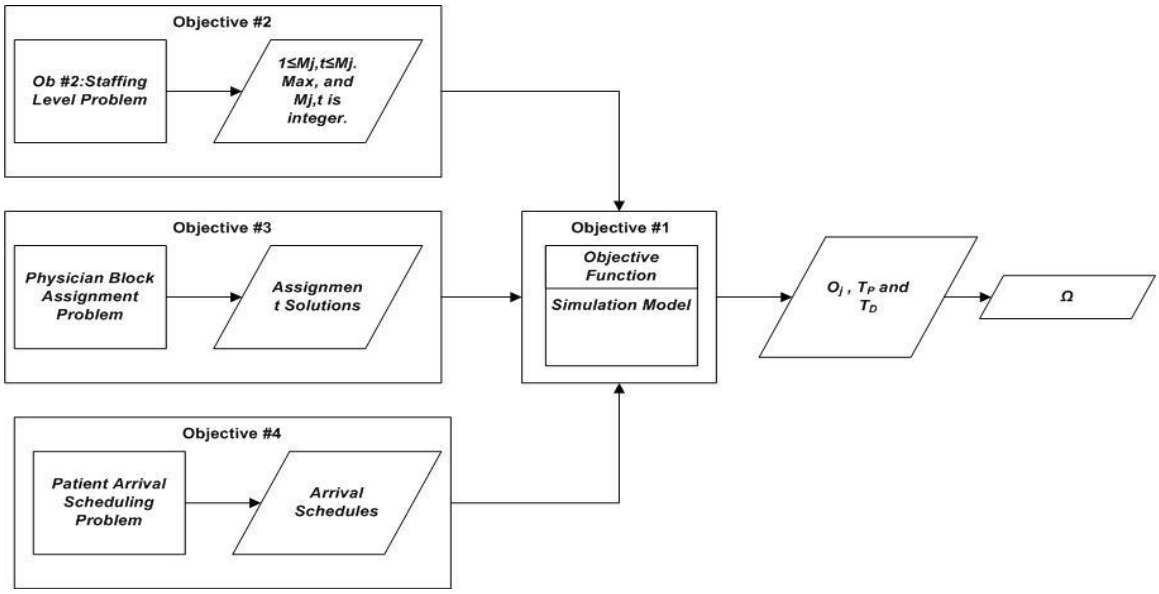


Figure 1.2 Connections among different research objectives

1.3 Research Significance

Compare to others' work, some significant research targets within different levels have been set up at the beginning or during process. The first target is about performance criteria settings which will be explained compared in the later sections. Multiple levels of performance criteria in this dissertation, which has been considered and expressed into one unite financial factor as different levels of measurements, are proved better single level. The advantages of this multiple levels criteria are making the optimal solutions more overall reasonable and with more inspects than single level criteria. Besides, cost factors are used to express different levels performance which can give administrators of the ambulatory surgical facilities direct investment options.

Next, three main tasks will have been illustrated in this dissertation containing staffing level optimization, physician group schedule optimization and individual patient scheduling by discrete–event simulation method quoting classical sequencing and scheduling rules. The staffing level strategy can help administer from ASFs clarify and save extra human resource cost under fixed physician schedules without sacrifice patients' satisfactory. The physicians' scheduling strategy has a significant role in ASFs because the ASFs are depending physician groups to assign them patients, and the scheduling efficiency matters because a better scheduling strategy would allow more patients to do the surgery within fixed time and improved the quality of care of ASFs because of the decreasing of physicians' delay. In the patient scheduling topic, a more patient-centered scheduling is offered by quoting classical sequencing and scheduling rules. It has reduced the time variations for patient arrivals and decreased the physicians'

delay time as well. In a word, ASF's total performance could be improved significantly by using the strategies prompted from this dissertation.

The remainder of the dissertation is organized as follows. In the next chapter the authors provide a brief review of the literatures in US healthcare and ASFs. The model construction and general assumptions are described in Chapter 3. Several experiments in staffing level optimization have been applied to the model and results and analysis are displayed in the chapter 4. Five different heuristic algorithms which would generate a daily physician group schedules have been explained and tested on ten environmental problems in Chapter 5, and one linear programming of balancing the resource and operative usage have been proved to dominant these heuristics which gave the best total performance. The statistical comparison among these results and the lower bound for the ten problems are also explained in Chapter 5. Chapter 6 describes the individual patient arrival scheduling problem with seven heuristic algorithms which are referred to classical flow shop problem have been introduced, and five algorithms' results have been compared with statistical analysis. Chapter 7 is about future work and reference is in Chapter 8.

CHAPTER 2

LITERATURE REVIEW

An overview picture has been sketched in this chapter over the development of global and US healthcare, appearance of ASFs, specific characters of ASFs, and current research work in the field. The first sub-topic (2.1) of this chapter is the historical view upon ASF in healthcare including healthcare improvement and IE applications (2.1.1) and appearance and specific characters of ASFs (2.1.2). The second topic (2.2) is focused on performance objectives composed by patient delay (2.2.1), physician delay (2.2.2), labor productivity (2.2.3) and facility utilization (2.2.4). In 2.3 modeling applications collect papers mainly in two categories mathematic programming (2.3.1) and simulation tools (2.3.2). Later in Chapter 4's staffing level optimization problems is more based on simulation results, but heuristic algorithms are added in Chapter 5's physician group scheduling problem. Standing as the most important part of ASF, operation rooms' analysis composed the whole section 2.4 on scheduling problems from facility (2.4.1), physician teams (2.4.2) and patients' perspectives (2.4.3). The last 2.5 is about Applicability of research, several examples from author' in this field have been demonstrated offered a way towards our case but some applicable problems have been prompted as well.

2.1 Introduction to ASF

As our nation struggles with how to improve the costly and troubled health care system, the ASF is a great example of a successful transformation in health care delivery. Ambulatory surgery, also known as outpatient surgery, is surgery that does not require an overnight hospital stay. Such surgery is commonly less complicated than that requiring hospitalization. The first facility was opened in Phoenix, Arizona, in 1970 by two physicians who saw an opportunity to establish a high quality, cost effective to inpatient hospital care for surgical services. Basically, (Cardoen, Demeulemeester et al. 2010) there are two major input patient classes in the literature review on operating room planning and scheduling, namely elective or non-elective patients and in patients or out patients. For the elective patients who will have a surgery appointment in advance, whereas the non-elective patients for whom a surgery is unexpected and hence needs to be performed urgently mostly on emergency rooms. On the other side, in patients refer to hospitalized patients who have to stay overnight, whereas outpatient typically enter and leave the hospital on the same day.

In 2006, (Karen A. Cullen, 2009 #59)an estimated 53.3 million procedures were performed during 34.7 million ambulatory surgery visits to 7000 different facilities in the US. Average times for surgical visits were higher for hospital-based centers than for visits to freestanding ambulatory surgery centers (147 minutes compared with 98 minutes).Frequently performed procedures included endoscopy of large intestine (5.7 million), endoscopy of the small intestine (3.5 million), extraction of lens (3.1 million), injection of agent into spinal canal (2.0 million), and insertion of prosthetic lens (2.6 million).

2.1.1 The Importance of IE Applications in Health Care

Human health has improved significantly in the last 50 years. In 1950 global life expectancy was 46 which rose to 61 years by 1980 and 67 years by 1998. However, in low and middle-income countries, where 80% of the world's population lives, malnutrition and infectious diseases account for significant numbers of premature deaths. Although high-income countries spend more on health than low-income countries, performance of health care systems varies markedly among them. France, which spends half as much as the U.S. on per capita annual health care, was ranked first in overall health systems performance (a recent report by the World Health Organization, health system performance includes not only measures of health, but also systems fairness and responsiveness.)In countries with no national health system, such as the U.S., a significant fraction of individuals have no health insurance coverage and thus have only limited access to health care.

The health care industry represents approximately 20% (recent data from times magazine) of the gross domestic product of United States currently and its expenditures are going to be doubled by 2050.(Gupta and Denton 2008) Moreover, health managers have to anticipate the increasing demand for surgical services caused by the aging population. In a word, there is no surprise that there is an increasing pressure for health care providers in efficiency and cost effective in health care services. There are many factors that affect the ability of health care's efficiency and effectiveness among the three basic cares: primary care, specialty care and elective surgical care.(Gupta and Denton 2008) The common big issues for the managers in these cases are how to maximize the labor productivity by using the least numbers of staff necessary to care for the patients

and how to schedule the patients and physician groups to facilities to get the maximize utilization under some fixed cost. The complexity is increasing from the primary care, specialty care to elective surgical care. Especially in the elective surgeries, there are uncertainties during every procedure time (pre operation time, surgery time and post operation time), in patients delay, in staffing or physician group delay. Upon these uncertain issues, the authors believe that a critical bottleneck lays with the application of Industrial Engineering & Operations Research models. Since 18th and 19th centuries, many people took time and efforts to apply science to process optimization in manufacturing and military systems. Nowadays, some successful IE applications have been used in airline, car rental agencies and hotels and the authors believe that IE/OR decision support techniques can be also applied in health care system to save budget and increase facility utilization at the same time.

Depending on the subspecialties involved, industrial engineering may also be known as, or overlap with, operations management and management science, depending on the viewpoint or motives of the user. For example, in health care, the engineers known as health management engineers or health systems engineers are, in essence, industrial engineers by another name. Basically speaking, there are a lot of fields in health care industry upon which engineers can work including health care financing, health care administration and regulation, health information technology and so on. As the information technology developed, more data (time, surveys and patients' records) tracked can bring us revolutionary efficiency improvement in this field. All these tracked factors have been helped in building the loop cycle in the system to get better

performance objects (patients' delay, doctors' delay, overtime control and facility utilization).

To conclude, along with the increasing occupancy and pressure nationwide, health care industry needs some urgent optimization methods to improve the performance objects. These complicated uncertainties in it can be viewed systematically and be solved through IE/OR applications, for example a lot of researches have been done in surgical suite optimization by simulations like (Cardoen, Demeulemeester et al. 2010) wrote generally; (Denton, Rahman et al.) offered a monte-carlo simulation model dealing with multi-OR surgical suite scheduling under different staffing scenarios;(Franklin Dexter 2002)illustrates the appointment scheduling problem for elective surgeries upon two patient-scheduling rules: earliest start time and late start time; while (Gul, Denton et al.) demonstrated DES(discrete event simulation) model to evaluate the appointment scheduling and also Genetic Algorithm to get near optimal sequences and appointment times, and another paper (Erdogan and Denton 2011)from Denton also considers situations when no patient show up, cancellation and dynamic cases from patients.

All of the above research is involved in the Ambulatory surgical scheduling problem which the authors will talk in details in the following sections. Besides, another Dexter's paper (Marcon and Dexter 2006) illustrated the scheduling sequence effects to PACU (post anesthesia care unit); (Alexopoulos, Goldsman et al.) described one tool concerned about children and poor people with low cost; (Carter 2010) did some research in scheduling in endoscopy suites according to physician average procedure time. Besides DES model, (Zhang, Murali et al. 2008) allocated the operating room capacity through MIP (mixed integer programming) and (Thor, Lundberg et al. 2007) concluded some

statistical process control in healthcare improvement. Though lots work have been done by IE applications, along with the increasing pressures from this filed and higher level of requirements from patients, further future work should consider more details in the system.

2.1.2. Specific Characteristics of ASF

The following Figure 2.2 (association) showing an decreasing trend in inpatient surgeries. (association) Avoiding hospitalization can result in cost savings to the party responsible for paying for the patient's health care. Frequently performed procedures included endoscopy of large intestine (5.7 million), endoscopy of the small intestine (3.5 million), extraction of lens (3.1 million), injection of agent into spinal canal (2.0 million), and insertion of prosthetic lens (2.6 million). The purpose of outpatient surgery is to keep hospital costs down, as well as saving the patient and physician group’s time. The Figure 2.1 below shows the ambulatory surgery’s application among different ages.

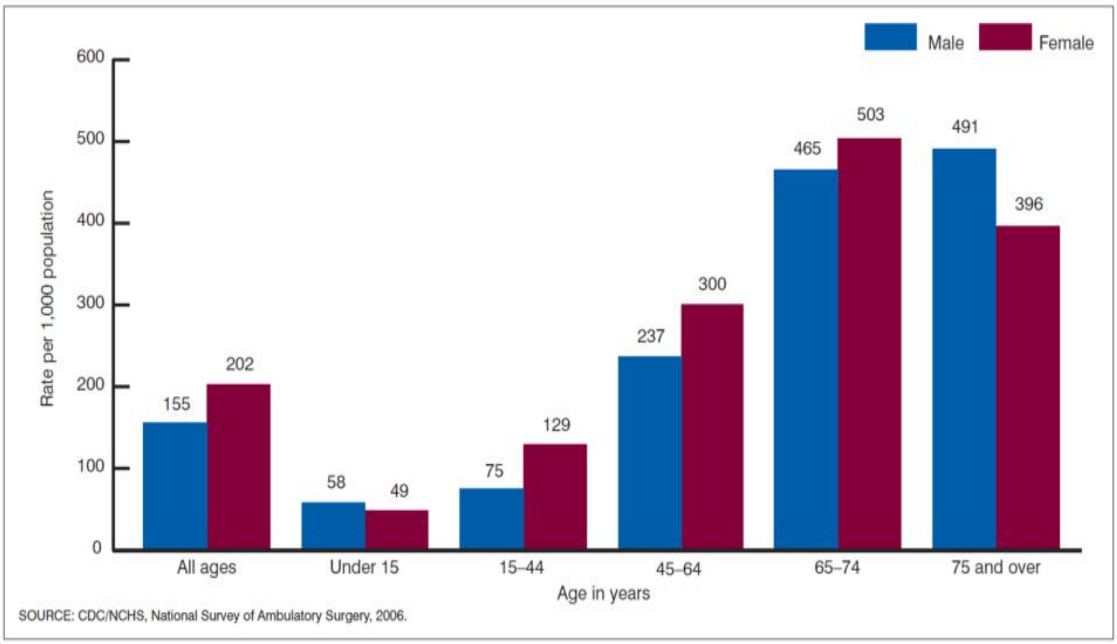


Figure 2.1 Ambulatory surgery's application among different ages.

Source: National Health Statistics Reports Number 11 January 28, 2009–Revised

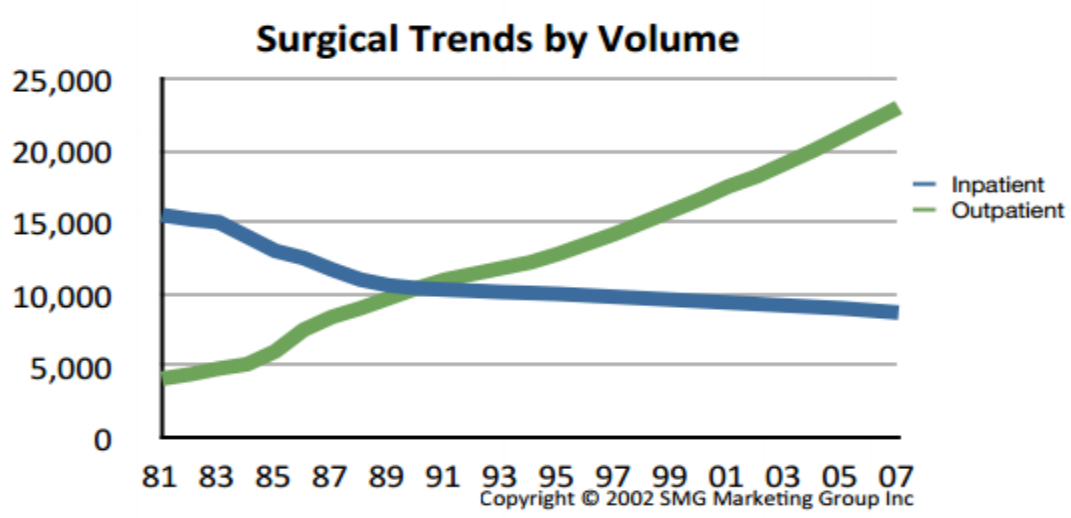


Figure 2.2 Surgical trend by volume.

Source: SMG Marketing Group INC

Ambulatory surgery centers (ASC) or ambulatory surgical facility (ASF), also known as outpatient surgery centers or same day surgery centers, was first built in Phoenix, Arizona. At that time, after faced scheduling delays, limited operating room availability and slow turnover times, two physicians got this opportunity to build the first ASF with more physician' involvement. Nowadays, some ASFs are still owned by physicians, the authors call it free-stand ASF, others are owned by hospitals. According to most recent data, 21% of ASFs' interests are owned by hospitals and around 3% ASFs are owned entirely by hospitals(association). The comparison between the HOPD and ASF in volume is shown in the following Figure 2.3 (Hair, Hussey et al. 2012) also compared the surgery time intervals, post-surgery time intervals and total time spent in freestanding ASF and hospital – based ASFs.

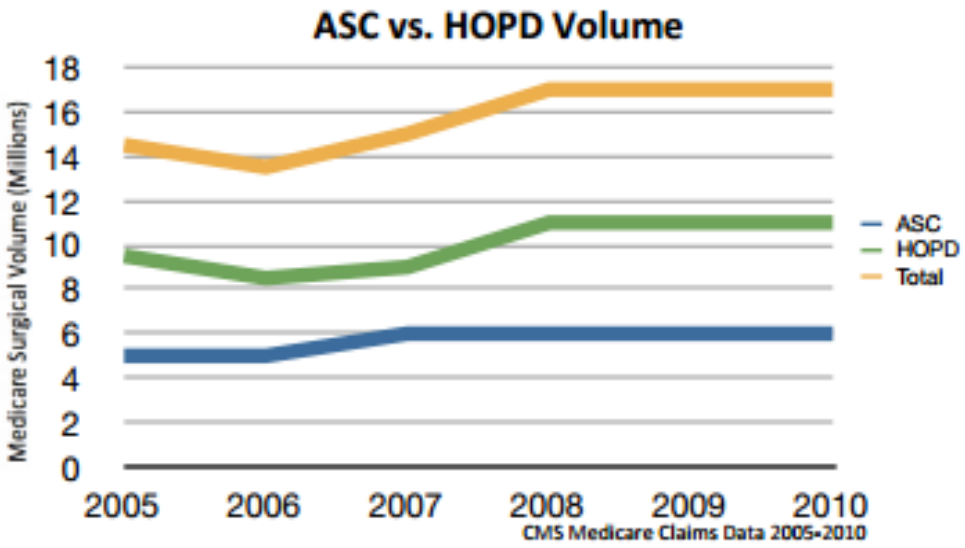


Figure 2.3 ASC vs. HOPD volume.

Source: SMG Marketing Group INC

ASFs now are adding considerable value (around $90 billion) to the US from some data from 2009, and ASFs employ the equivalent of approximate 117,700 full-time workers. Accordingly, a lot of researches have been done in this field. (Durant 1993) and (Durant and Battaglia 1993) are two early papers gave us a view in future ASF's development directions and some government policies. (Roberts 1994) is a later paper suggests that some newly formed centers build the accreditation systems.(Reis, Mosimann et al. 1999)recommend implementing ambulatory surgery in a teaching hospital and encourage the expansion of this practice. (Joshi and Twersky 2000) introduced and highly suggest a new paradigm "fast tracking" which involves transferring patients from the operating room to the recovery unit prevent complications. Besides physical equipment, (Yeung, Cheung et al. 2002) and (Franklin Dexter and Margaret Hopwood) both add some surveys to get feedbacks from the patients to continuously improve the total performance. In 2010, there are two papers concerned different parts of ASFs,(Hollingsworth, Krein et al. 2010) evaluated how opening of an ASF center impacts stone surgery use in a health care market and assessed the effect of its opening on the patient mix at nearby hospitals; the author just stayed in one hospital which converted to an ASF, which gives him a chance to get a comparison the time intervals between a hospital and an ASF.

2.2 Performance Objectives

There are various research aspects can be set as performance measures, from the paper (Cardoen, Demeulemeester et al. 2010), they concluded the basic eight objects from

previous researchers: waiting time, throughput, utilization, leveling, makespan, patient deferrals, financial measures and preferences. one of the performance measures in the paper (Weng and Houshmand) is to maximize the patient throughput. In the paper (Weng and Houshmand), they used throughput, time in system and queue times and lengths with total cash flow to get alternatives for resource or scheduling requirements for a local clinic. The following section is arranged to overview papers in patient delay, physician delay, staffing utilization and facility utilization.

2.2.1. Patient Delay

Generally speaking, there are various criteria are proposed to evaluate the performance of the planning and scheduling methods. Among the eight objects, the most common complaints from patients are the waiting time, and it is also an important part of patients' satisfactory. Especially in some emergency cases, this waiting time can be critical because it relates people' life. However, how to define the waiting time can be different in different cases. In the paper (Franklin Dexter and Margaret Hopwood), since it's a block scheduling problem for the operating room, the author defined two different types of patient waiting time: indirect waiting time and direct waiting time, and the latter concept is what the authors usually define patient waiting time. And for the first indirect waiting time, is the time starts when the patient submit his/her willing time windows till receive the confirmation time. As the common sense, when the authors collect more information of patients, it getting better scheduling results for hospitals, however, it accompanied with cancelations during the waiting period as well. However, this search is also based on one survey about patients' preferences for surgical waiting time (2-4 weeks are acceptable).

In the mapped arrival process of review paper (Gupta and Denton 2008), the author explained the process types in different situation. The single batch process with irrelevant inter-arrival times is commonly assumed in elective surgery start times; the unit process which assumed to at a time and at random time epochs is commonly used in primary and specialty care appointment scheduling design; the periodic process happened when all requests are accumulated at the end of one period which is commonly assumed in specialty and elective surgery cases and then the single batch process can be treated as one of this category with intervals covering entire booking horizon, but the former one's model is quite distinct from this.

In addition to these different types of patient waiting time categories, some papers which concerned about patient scheduling are taken patient delay as the performance measures. (Gul, Denton et al.) took both DES (Discrete Event Simulation) and GA (Genetic Algorithm) to find the optimal scheduling strategies for patients with patient waiting time and overtime as objects. In the paper(Hsu, de Matta et al. 2003) formulate the patient scheduling problem as variants of the no-wait time when to get minimized number of nurses at post anesthesia care unit. There is another object, throughput, is closely related to patient waiting time, and lots of papers have been involved with this object.

2.2.2. Physician Delay

As one of the most expensive resources in operating rooms, physician groups' satisfactory becomes an important object for hospitals. Although from the paper (Cardoen, Demeulemeester et al. 2010), the lists of performances' measure table, more papers focused on patient delay, in ASF and other physician group based systems, the

physician delay is the most important factor. In the ASFs, the manager would try best to have physician groups' satisfactory because it's the physicians who bring them patients. However, this satisfactory mainly comes from the less waiting time during the surgery process. In another Denton's paper (Denton, Viapiano et al. 2006), physician's delay also defined as operating room idling time, and they studied how the sequencing affects patient waiting time, physician idling time and operating room overtime. Accordingly to reduce the physician idle time, more physician blocks' scheduling problems have been the subject of recent research and I will talk about this in detail in the later sections.

2.2.3. Labor Productivity

As another big issue for managers to consider is the labor productivity. (Franklin Dexter and Margaret Hopwood)talked about this issue in OR managers side, they said that the OR managers must try to maximize "labor productivity" by using the least number of staff necessary to care for the patients for the first step task. When the authors have more staffing members, it will increase the total operating budget for hospitals, on the opposite side, not enough staffing members (When only concern about staffing members like nurses and medical assistants) will decrease the satisfactory from both patients and physician groups by increasing their waiting time. For the time they work are divided into two basic parts: regular time and overflow time. (Cardoen, Demeulemeester et al. 2010)Utilization (here is also productivity) actually refers to the workload of a resource, whereas under time or overtime includes some timing aspect. On the one hand, setting the overflow payment for staffing members is realistic and necessary to get patient through, on the other hand, it also reflects the regular utilization of labor resource and controls the idle resource waste as well.

Some hospitals may hire some lower levels of nurses to be on the overflow time shift, while others still keep the original nurses on duty but with some extra payments for that. From hospitals' financial side, it is better not have over time for nurses or medical assistants, and also from the ergonomic side, longer high-concentrate working hours may lead to fatigue quickly and also result to unsafe factors. In one of Dexter's paper (Marcon and Dexter 2006) they analyzed the impact of sequencing rules on the phase I PACU(post anesthesia care unit) staffing and over-utilized operating room time resulting from delays in PACU admission, and they suggested some adjustment in PACU nurse staffing around the times of OR admissions.

In the model, the authors also have patients' waiting time and doctors' delay time as two of the performance criteria. As it mentioned in the literature review, there are other criteria can be used in different models like the patient through put number. In particular, patient throughput numbers can be changed by the capacity of facilities and the operating schedule. Since if the authors just ignore other changing parameters would lead to inefficiency output data for the key focus, the authors have some assumptions ahead before the experiment, and one of them is make the difference between patient through put number and total patient coming number less than 5 units. In addition, the time load and numbers for operating rooms are fixed elements since the authors focus on labor utilization. After these assumptions about other changing factors, the objective, best labor resource utilization can be gained with best optimal time and cost balance and without other conflict factors.

2.2.4. Facility Utilization

Besides the two waiting time criteria below, the authors also have the other criteria from the resource utilization side. On the one hand, utilization should be maximized as underutilization operating rooms represent unnecessary costs. On the other hand, operating rooms without any time buffers could easily result to labor's overtime cost and other uncertainty costs. Although in the case, the facility cost is fixed in a head of time, many studies elaborate on this trade-off and evaluate procedures based on the OR efficiency. Not only in OR, but also in ICU(intensive care unit), there is one paper (Zhu 2009)focused the ICU beds because lack of it may cause ambulance diversion and surgery cancellation, DES is used and real data from the hospitals are as inputs and finally they offer better solutions to trade off the utilization of ICU beds and waste of resources.

2.3 Modeling Applications

In an overview, industrial engineering typically use computer simulation (especially discrete event simulation), along with extensive mathematical tools and modeling and computational methods for system analysis, evaluation and optimization. As listed in the review paper about operating room planning and scheduling (Cardoen, Demeulemeester et al. 2010) table 7 solution technique, there are mathematic programming includes :LP (linear programming), quadratic programming, Goal programming, MIP (mixed integer programming), dynamic programming, column generation, branch-and-price programming and so on. The two common simulations are DES and monte-carlo. After

some simulations, some heuristics are generated which some has improved ones called Meta-heuristic with simulation annealing, tabu search, GA (genetic algorithm) and others. Some papers involved more than one methods in the research like the paper (Gul, Denton et al.) DES and GA at the same time; (Lamiri, Grimaud et al. 2009) combined Monte- Carlo simulation to optimize the surgery planning when OR rooms are shared by elective and emergency case at the same time.

2.3.1 Mathematic Programming

LP (Linear programming) is a mathematical method for determining a way to achieve the maximum or the lowest cost in a given mathematical model for some list of requirements represented as linear relationships.(from Wikipedia) (Erdogan and Denton 2011) formulated with stochastic LP formulations in the appointment scheduling problems of patients fail to show up in the first model and another model with multistage LP program to solve dynamical customers' request.

As one subject of LP, MIP (Mixed Integer Programming) is mostly used in discrete optimization problem such as transportation, airline crew scheduling and production planning. In the paper (Zhang, Murali et al. 2008), they developed a finite-horizon MIP model for allocating operating room capacity to specialties. A tabu search-based heuristics algorithm is generated in the paper (Hsu, de Matta et al. 2003) to get minimized nurses numbers in PACU.

2.3.2 Simulation Tools

Simulation is a tool in which a mathematical is built to act like a system of interest and it is popular in engineering and management sciences for analyzing problems which there is

uncertainty. The DES (discrete- event simulation) method is widely used in health care system because of the variability and complexity of the process within health care system. (Ferreira, Coelli et al. 2008)Discrete –event simulation is a computer modeling strategy in which events are assumed to take place one at a time, with subsequent events happening exclusively after the end of the predecessor. Discrete Systems– the state of the system changes only at discrete points in time due to the occurrence of certain events. Whereas Continuous 6me systems the state changes continually. In engineering, discrete-event models are commonly used to study the behavior of systems, their performances, limits and future states. In our simulation case, since it involved a lot of processes and different satisfactory levels, the authors build the model using discrete event simulation.

(Alexopoulos, Goldsman et al.) is one paper using DES to build one flexible simulation serve for small facilities for the poor which has problems in finances and personnel.(Ferreira, Coelli et al. 2008) also used DES but to optimize the patient flow in a large hospital surgical center. (Marcon and Dexter 2006) applied DES to show the importance of nurse capacity in the PACU. By using DES, wullink compared two approaches to deal with emergency surgery by have some ORs reserved or sharing with elective patients. Some common DES software like

- Arena - a simulation and automation software developed by Rockwell Automation. It uses the SIMAN processor and simulation language.
- Flexsim – FlexSim Healthcare includes a whole library of objects that are ready out-of-the-box for building almost any healthcare model
- Simio -Models built with all four Editions are fully compatible both up and down the product family. All four products provide the same powerful 3D object-based modeling environment.

Monte Carlo simulations are a class of computational algorithms that rely on repeated random sampling to compute their results. Monte Carlo methods are often used in computer simulations of physical and mathematical systems. These methods are most suited to calculation by a computer and tend to be used when it is infeasible to compute an exact result with a deterministic algorithm. This method is also used to complement theoretical derivations. (Lamiri, Grimaud et al. 2009) combined Monte- Carlo simulation to optimize the surgery planning when OR rooms are shared by elective and emergency case at the same time. This method is also used in paper to find the functional ICU beds under conditions when the operative procedures were canceled by Monte Carlo Simulation.

2.4 OR Room Scheduling Modeling

The single largest cost from hospitals' surgical delivery comes from the OR room because of the salaries for OR staffs and nurses. Operating rooms (ORs) have been estimated to account for more than 40% of a hospital's total revenues and a similarly large proportion of their total expenses, which makes them a hospital's largest cost center as well as its greatest revenue source. ORs represent the hospital's greatest revenue source (Denton et al., 2007) For example, the French health ministry and health regulators have encouraged OR managers to achieve 80% or more OR utilization. Therefore, the first task for the OR managers is to reduce the least necessary staff and nurses members. Then since the appointment scheduling problem just lies at the intersection of the efficiency and timely access to health services, which would be the

second important factor the managers will concern after the resource capacity. According to different main research objects in the system, the OR scheduling can be divided into physician block scheduling and patient scheduling. Though OR scheduling comes the most important part in surgical care, it will affect the PACU scheduling in which is another topic a lot of people have done researches and it will be affected by other connection factors such as patients' cancellation and no show up rate .

Similar in some degree in paper (Cardoen, Demeulemeester et al. 2010) decision delineation: they distinguish these decisions between the discipline, the surgeon and the patient level. The discipline level they defined as unites contributions in which decisions are taken for a medical or department as a whole such as operating room time. While the surgeon level will arrange the operation room, time and time block for the surgeon. Besides, in patients' level, they are usually divided into elective and non-elective categories. In addition, there are other people did optimizations in ICU (Intensive Care Unit) room and the authors will explain them in details in the following paragraphs.

2.4.1. Resource Capacity Analysis

In some papers, they define a master surgery schedule as a schedule which specifies the number and the types of operating rooms, the hours that operating rooms are available. In the paper (Weng and Houshmand),by maximizing patient throughput and minimizing patient flow time, they found the 6 second year residents and 2 medical assistants is the optimal staff size in the local clinic. Partly, the paper (Marcon and Dexter 2006)shows the importance of enough nurses in the PACU through the DES. In the paper by Philip about ICU beds, the authors found out the functional ICU capacity.

2.4.2. Physician Block Scheduling Problem

From one of Franklin Dexter's papers (Franklin Dexter and Margaret Hopwood), the authors tried to determine the appropriate amount of block time to allocate to surgeons and selecting the days on which to schedule elective cases can maximize operating room scheduling. They also defined the OR utilization: equals the time and OR is used divided by the length of time an OR is available and staffed. To get maximized OR utilization, several algorithms are generated such as next fit, first fit, best fit and worst fit and next fit produced OR utilization values as high as the other algorithm and it's the simplest. To conclude, they found out the most importance parameter affecting OR utilization is the mean length of time patients have to wait before surgery: the longer patients have to wait, the less unused block time there will be. According to some survey data, the patients are provided that open block time within 4 weeks; otherwise, they will be scheduled in overflow time outside the block time.

Sequencing and scheduling is raised by scary resources' allocation to activities through the time in production planning, computation control and other general situations. The three main topics included are single or parallel machine sequencing, flow shop sequencing and job shop scheduling. Since the definition of scheduling almost covered sequencing, though they focused on different aspects, the scheduling is chosen to stand for sequencing and scheduling in the following content. Single-machine scheduling or single-resource scheduling is the process of assigning a group of tasks to a single machine or resource. The tasks are arranged so that one or many performance measures may be optimized. Parallel machines are parallel identical machines meaning that tasks or jobs can be finished by either of the machines. The main difference between

single machine sequencing and flow shop sequencing is that more machine quantities and given process order (the definition of flow shop scheduling is given later). However, the range of job shop scheduling is wider than that of flow shop scheduling, for example, both with process orders, usually one job is not allowed to rework in the same machine in the flow shop scheduling problems but there is no path route rule for jobs in job shop scheduling problems.

With about 70 years' investigation, major findings include: Graham had already provided the List scheduling algorithm in 1966, which is (2 − 1/m)-competitive, where m is the number of machines. Also, it was proved that List scheduling is optimum online algorithm for 2 and 3 machines. The Coffman–Graham algorithm (1972) for uniform-length jobs is also optimum for two machines, and is (2 − 2/m)-competitive. In 1992, Bartal, Fiat, Karloff and Vohra presented an algorithm that is 1.986 competitive. A 1.945-competitive algorithm was presented by Karger, Philips and Torng in 1994. In 1992, Albers provided a different algorithm that is 1.923-competitive. Currently, the best known result is an algorithm given by Fleischer and Wahl, which achieves a competitive ratio of 1.9201. A lower bound of 1.852 was presented by Albers. Taillard instances has an important role in developing job shop scheduling with makespan objective.In 1976 Garey provided a proof that this problem is NP-complete for m>2, that is, no optimal solution can be computed in polynomial time for three or more machines (unless P=NP).

2.4.3. Patient Scheduling Problem

Standing at the patients' side, there are basically two types of patients: non-elective (emergency cases) and elective cases. The patients' types of that hospital depend on the hospitals' type. If it is free stand ASF, it would only have elective patients. If the ASF

was combined with ER (emergency room), both types may have to share some resources. Only few studies have considered the situation where ORs are used to provide service to elective and emergency. (Gerchak et al), proposed a stochastic dynamic programming model for the advance scheduling of elective patients for ORs serving elective and emergency patients. They focus on how many additional requests from patients assigned for that day. Some people also compare strategies between having reserved beds for emergency cases and sharing with elective cases like in the paper by Wullink et al, (2007).

2.4.4. Others

Although a lot of efforts are put in the ORs, efforts to increase OR utilization can affect the functioning and the efficiency of other stages of the surgical process such as the phase I post anesthesia care unit (PACU). The paper from Marcon and Dexter (Marcon and Dexter 2006)analyzed the impact of sequencing rules on the PACU staffing and over-utilized operating room time resulting from delays in PACU admissions. Seven sequencing rules are tested: random, LCF (Longest Cases First), SCF (Shortest Cases First), Johnson, HIHD, HDHI and MIX, and the best rule is HIHD (Half Increase in OR time and half decrease in OR time) which can offer smooth patients' entering. On the other side, they against the LCF rule which will generate more over-utilized OR time and require more nurses in PACU. Others researches also have been done in the ICU about the number of occupied ICU beds at which operative procedures were canceled if they were known to require an ICU stay.

2.5. Applicability of Research

This section is focused on some research results which have been offered by previous authors, the challenges of applying these into real-world have also explained after some examples of results. Figure 2.4 illustrates the empirical probability in OR1. Distributions for two specific examples of surgeries in OR1.The structure of these example distributions is typical of uncertainty in surgery durations, where there is a fairly significant mass of probability confined to a predictable range, and a tail indicating a lower probability of extended surgery duration resulting from unexpected complications. Instances of the stochastic linear programming model were created using 10, 000 scenarios. To evaluate the effect of sample size 100 replications of the optimal solution with K = 10, 000 were performed for each of the 5 daily schedules for the OR1 weekly schedule. The confidence intervals for the optimal solution for OR1 test models ranged from approximately ±1 to 2.5% relative to the mean. Based on these results the authors use K = 10, 000 scenarios for the remainder of the numerical experiments.

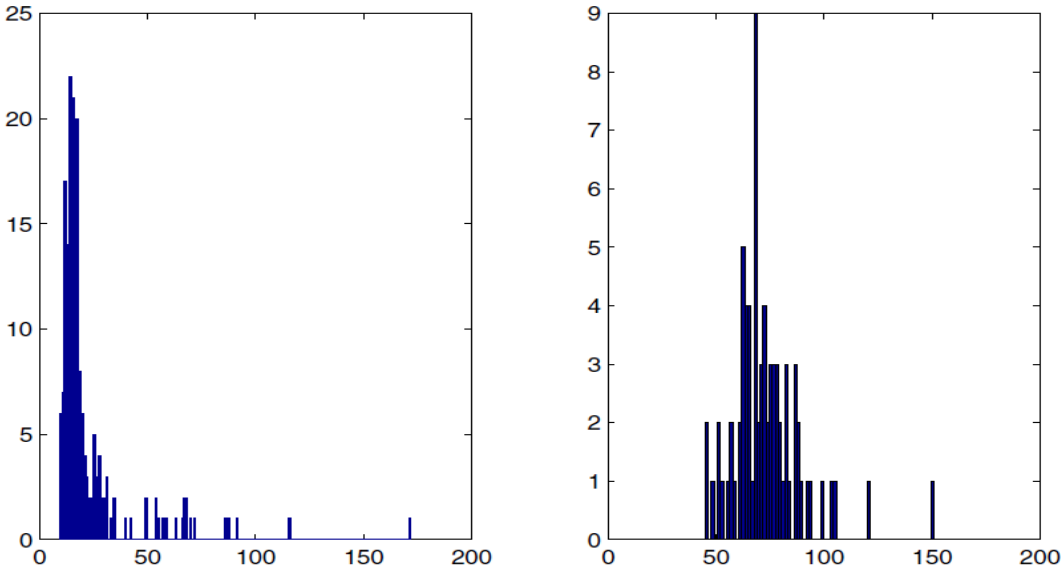


Figure 2.4 Process duration (minutes) distributions for two surgery types.

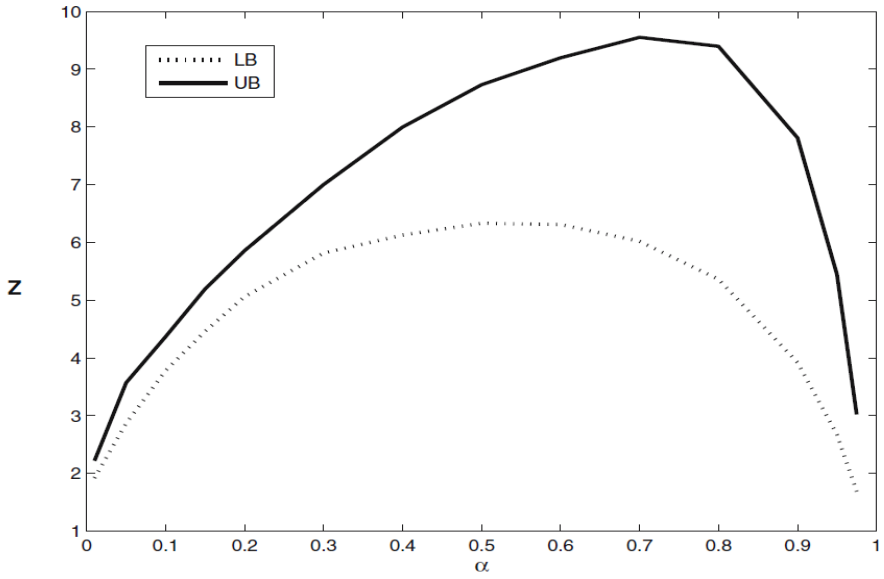


Figure 2.5 Range in the optimal objective function for the best and worst sequences of surgeries as a function of the relative difference in the waiting cost coefficient.

After the formulations, simulations tools have been used, even a theoretical strategy has been prompted at last, many researches provide that a thorough testing phase cannot simply implemented in practice and it is hard to find statements in contributions that explicitly confirm the implementation and use of the procedures in practice. Though most of the research data is from real hospitals, only limited research is performed to indicate what planning and scheduling expertise is currently in use in hospitals. In the review paper (Cardoen, Demeulemeester et al. 2010) also demonstrated that it is hard to provide details on the process of implementation and they encourage the provision of additional information on the behavioral factors that coincide with the actual implementation under some implementation can be assumed in some situations. Therefore, in the future, more work can be done in verifying the research results.

CHAPTER 3

SIMULATION MODEL OF AN AMBULATORY SURGICAL FACILITY

Simulation modeling is a powerful analytical tool for modeling the behavior of systems with complex processes and multiple sources of variability. As such they are ideally suitable for the study and analysis of healthcare systems. The review indicates that simulation is also and effective and amenable tool for the study of ASFs. The authors find that it is difficult if not impossible to develop exact analytical models for the following reasons: (a) Time related uncertainties - Three system uncertainties characterize the problem (ii) Surgery time variance (ii) Physician arrival delay and (iii) Patient arrival delay; (b) Resource Capture Complexities – Patient flows vary significantly and capture/utilize both staffing and/or physical resources at different points and varying levels; and (c) Processing Time Differences – Patient care activities and surgical operation times vary by type and have a high level of variance between patient acuity within the same surgery type. In this chapter we present the development of an ARENA based simulation model that accurately characterizes the activities and operating behavior of a typical ASF.

This chapter is organized as follows. Section 3.1 defines all the resources in the model which will be used in later model constructions; The second section (3.2) of this chapter is about model constructions in details to a general ASF including: ASF operating process analysis with assumptions and an event flow chart clarify logic connections between process; (3.3) introduces model input data in different tables under one scenario

and general performance objective function set up is in (3.4); (3.5) states that rational reason of choosing discrete-event simulation(3.5.1) and key surgical processes converting into Arena model (3.5.2); statistical validation of the simulation model is in (3.6); and the last section of the chapter (3.7) concluded causes of uncertainties in ASF system and those changeable decisions which the authors could make to optimize ASFs. The listed topics in the conclusion will be analyzed in details in following chapters.

3.1. The Model Building Approach

A wide range of professional group, insurance industry and federal healthcare practices and regulations govern specific clinical procedures. In contrast patient flows and resource use behavior are less standardized and tend to vary between different healthcare facilities across the USA. For ASFs also operational systems do vary but generally are less variant when compared to patient flows in hospitals or large clinics. The first step of the model building approach was to therefore research the current operational flows of ASFs, with the goal of identifying a typical operation process flow. Specifically the following activities were carried out:

- *Direct Work Study* – The authors were provided access to four different ASF facilities all located in New Jersey. For some extensive access was provided with multiple days of visits recorded (Meadowlands Outpatient Surgery), while for other limited access was provided with a single day of access but access to operations records and time sheets (Virtua Healthcare). Note that HIPAA regulations and standard practice limit the flexibility that

ASF managers have in providing access. Most common work study information source was in the nurse break room where the staff were able to allocate time to the team. All team members had to wear nursing attire within the ASF clean areas. For the different observation data was recorded on an Excel template designed for this purpose.

- *Discussions/Interviews with ASF Staff* – The authors were able to have discussions/interviews with 8 staff members at different ASF facilities all located in New Jersey. An example facility is Surgicare of Central Jersey which specializes in Gastroenterology, Orthopedic and Opthalmic surgeries. This activity occurred after the direct works study, and many of the questions were focused on clarifying issues identified in that activity.
- *Review of ASF Operations as Reported in the Literature* – Several reports identify and describe different parts and operations with an ASF. A key source of such data is the Ambulatory Surgery Center Association (ASCA). The review activity was used to validate the constructive assumptions and modify the ASF process flow model developed in this chapter.

Based on the above activities the authors have identified (i) the significant cost variable resources in an ASF, (ii) the associated patient flow process in a typical ASF and (iii) the performance relationship between resources and patient flows. Further, all needed logical rules and data needed to construct the model were developed.

3.2. Resources in an Ambulatory Surgical Facility

From the work flow analysis four primary categories of resources are identified: (i) Staffing or labor resources need to run the ASF healthcare activities these are estimated to be 32% of operating costs nationally (ii) Administrative resources consisting mainly of administrative staff needed to run the non-healthcare activities at the ASF, estimated at 20% of costs, (iii) Medical and Surgical Facilities/Equipment needed to provide the needed quality of care and surgical support (e.g; preoperative beds, operational bed and postoperative beds), estimated at 29% of costs, and (iv) Physicians who perform the surgery including anesthesiologists or other professionals that are directly associated with the physician, this resource category is a not a cost resource for the ASF since they are directly compensated by the insurance company. The authors found that the first category is the only real variable or controllable cost for an ASF, and we identify several different sub-categories that are modeled here.

Staffing Resource - Nurses: ASF nursing is characterized by rapid and focused assessments of patients, and building of immediate patient relationships. These nurses work in outpatient settings, responding to high volumes of patients in short term spans while dealing with issues that are not always predictable. On the basis of different medical care jobs, education background and skills, different levels of nurses are categorized. A licensed practical nurse (LPN) typically handles preoperative and post-operative care, including starting IVs, assisting patients with bathing and dressing, and providing bedside care during recovery. In the operating room, registered nurses (RNs) or advanced practice nurses assist the surgical team and coordinate all room activity. Surgical nurses are also

responsible for educating patients on procedures prior to surgery, adjusting treatment plans, and teaching them about post-operative self-care.

The approach followed here is to setup two levels of nurses, group A and group B, which together can accommodate most patient flows. “B” group is a group of mixed nurses which are composed by more LPNs than RNs like nurse anesthetists; while “A” group nurses are in more advanced skill level who will assist physician groups in surgery process like first assistants, surgical nurses, surgical technologists and operating department practitioners.

Staffing Resource – Medical/Tech Assistants: These assistants perform a range of clinical and healthcare technology tasks to support the work of physicians and nurses. They perform routine tasks and procedures such as measuring patients' vital signs, administering medications and injections, recording information in medical records-keeping systems, preparing and handling medical instruments and supplies, and collecting and preparing specimens of bodily fluids and tissues for laboratory testing. For preparation of some less complex surgeries, medical/tech assistants will in charge of preoperative and postoperative processes. The approach here is to model only one group of assistants.

Physician Resource: Physicians is the key resource in an ASF and no surgery can be performed without them. ASF physicians specialize in a specialty and hence perform a specific sub-group of ambulatory surgeries (e.g. gastroenterology). Physicians are not a direct expense resource for an ASF, in that

they are on the ASF payroll. It is common industry practice for physicians to be organized into groups. The authors assume that all physicians in a group can perform all surgeries in associated specialty. An ASF will typically have several physician groups working there, as a result the ASF can handle a wide range of surgeries.

Facility Resource: Common facilities in an ASF include lounge or registration area, preoperative beds (PreOP), surgery operating rooms (Surgery OR), and postoperative beds (Post OP) or post anesthesia care unit (PACU). Lounge room in ASFs is a place for patients waiting for the preoperative process. In lots of hospitals' introduction webpage, either beds or rooms numbers have been used to earn patients' surgery confidence upon their facility capability. To have the same unit to measure in preoperative, surgical and postoperative spots, beds are chosen instead of rooms to describe the ASFs' facility situation. Preoperative: The preoperative phase is used to perform tests, attempt to limit preoperational anxiety and may include the preoperative fasting. It starts when any preoperative bed and staffing resource are available, otherwise patients should wait in the ASF lounge room, and ends when patients are transferred to operative room.

The intra-operative period begins when the patient is transferred to the operating room bed and ends with the transfer of a patient to the PACU. During this period the patient is monitored, anesthetized, prepped, and draped, and the operation is performed. It starts when any operative beds and needed resources are available, and ends till patients are sent to PACU. Some clean up time should be left after the surgery. The postoperative

period begins after the transfer to the PACU bed and terminates with the resolution of the surgical squeal. It is quite common for this period to end outside of the care of the surgical team but the postoperative time the authors tracked is only in ASFs.

In the model, several surgeries are performed in some specific ASFs and parts of them are in a less complexity level. Those surgeries will be assisted mainly by Medical Assistant Group, which is composed mostly by medical assistants. Here, medical assistants and nurses are called staffing resources.

3.3. ASF Process Flow Model

A generalized ASF process flow model was developed from the work flow analysis reported earlier. This flow model is representative of the operations seen at most ASFs. Figure 3.1 summarizes the macro flow while Figure 3.2 shows a flow chart including some of the key logical decisions associated with patient transfers.

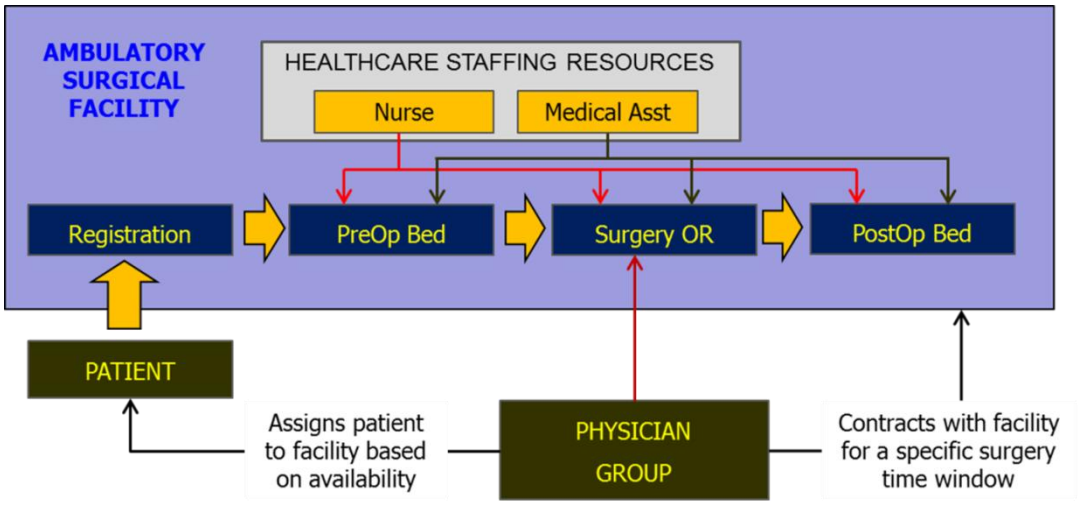


Figure 3.1 Macro flow model of an ASF.

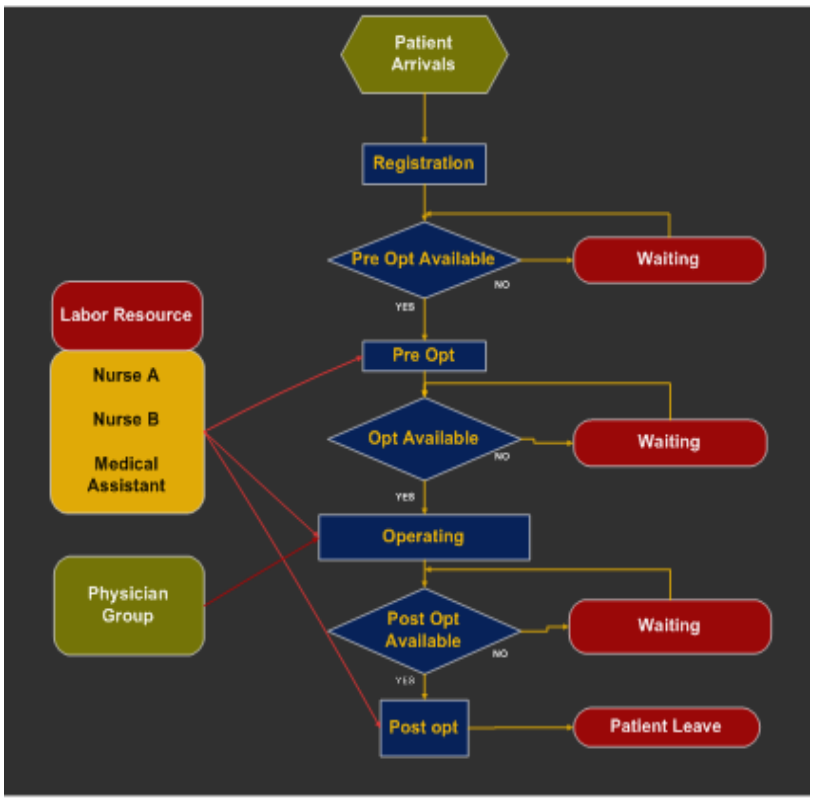


Figure 3.2 Patient transfer logical flow model of an ASF.

As shown in figure 3.1, the resources under the direct control of the ASF include two staffing resources and four facility resources. All arriving patients will flow sequentially through the four resources. Registration only uses administrative resources that are fixed to the registration desk and hence the staff resource is not independently modeled. Based on the work flow investigations the patient view flow process is described by the following steps:

1. Patients are scheduled to arrive at the ASF at a given time, on a given date, for a specific surgery to be performed by a specific physician group. *Assumption* – patient arrivals are uncertain and are described by a Poisson arrival process, and arrival sequence follows schedule sequence.

2. On arrival patients enter a common queue for the registration desk, where there is one or more administrative staff. *Assumption* – registration time is normally distributed with an average time of 5 minutes and standard deviation of 0.5 minutes based on the work flow survey.

3. Patient wait in the lounge area until a PreOp Bed is available, at which time they are moved to the PreOp Bed. *Assumption* – There is a fixed 10 minute setup time for each PreOp bed between patients, this does not require any of the modeled staffing resources. All beds have multi functions and not linked to a specific procedure and can therefore be used for any type of patient.

4. The PreOP procedure is approximated by three different types each of which uses different staffing resources. Additionally three time length distributions are possible each with a different mean process time. The specific type of PreOP is dependent on the surgery type and the patient acuity. Patient waits in the PreOP bed for the procedure to begin until the needed staffing resource is free and captured for the entire procedure time. *Assumption* – For each surgery two patient types based on acuity are modeled, the patient type will determine the PreOP type. Patient type is known prior to arrival. PreOP length is determined in real time after patient enter the PreOP, actual time follows a triangular distribution.

5. Patient remains in PreOP (blocked) until a Surgery OR is available. Additionally the move only occurs if the number of patients waiting in Surgery OR for a physician group is less than the physicians in that group. For example, if there is only physician in the group and there is already one patient waiting in Surgery OR for this physician

then the patient remains in PreOP. *Assumption* – Patient transfer time is zero, and all staffing resources are released immediately after PreOP procedure end.

6. Patients wait in the Surgery OR for the procedure to begin until the needed staffing resource and physicians are free and captured for the entire procedure time. Capture only occurs when all resources are available. Each surgery has two levels with different staffing resources, but the process time is the same. *Assumption* – Patient transfer time is zero, and all staffing resources are released immediately after surgery end. Surgery level is determined by patient type and surgery time follows a truncated normal distribution.

7. Patient remains in Surgery OR (blocked) until a Post OP bed is available. The PostOp process similar to PreOP has three time length distributions. The PostOp process uses the required staffing resources for only short periods in the start and end of the process. After patients have been transferred to the post operation rooms, staffing members are only needed in the first and last 10 minutes other than companying during the whole recovery process. In these total 20 minutes, nurses and medical assistants can only serve one patient at a time. The process can only start and end therefore when the resource is captured for these intervals. *Assumption* – PostOp bed is released immediately after process end and there is no blocking. There is a fixed 5 minute setup time for each PreOp bed.

8. Patient exits the ASF after release from the PostOP.

From the above flow process the authors know that staffing members will involve in the whole process from PreOP to PostOP, while the physician group members are in need of showing up only during surgery process after all preparation finished. Like what

the authors mentioned in the assumptions, all patients can be served only after beds are available for that process, otherwise, they should wait in queue for the bed (shows in a condition decision box).

3.4. Physician Schedules and Patient Relationships

The ASF work flow provided detailed insights on both physician scheduling arrangements at ASFs and the patient relationships. Patients are associated with a physician group, and the facility is therefore fully dependent on the physicians groups for the surgical business. Here the authors first introduce the setup of the physician schedule which relates the operations of the ASF to the different physician groups who perform surgeries at the facility. Most ASFs operate on a 9 or 12 hour day, which is further divided into 3 schedule blocks. A schedule block is defined as a continuous window, usually 3 to 4 hours long, during which assigned physician groups can schedule their surgery patients. Physician groups will contract with the ASF for to perform surgeries during one or more blocks. Clearly, these contracts must be within the capacity constraints of the ASF. In this research, it has been structured this relationship into the scheduling matrix shown in Figure 3.3. Note that some ASFs are not well organized and the physician scheduling arrangements tend to be more loosely setup.

Physician Group	Number of Physicians	Scheduling Blocks		
		8 am to 12 Noon (Morning)	12 Noon to 4 pm (Mid Day)	4 pm to 8 pm (Evening)
#1 - Gastroenterology	3		⊗	⊗
#2 - Orthopaedics	2	⊗		
#3 - Gastroenterology	1	⊗	⊗	⊗
#4 - Opthalmic	2		⊗	⊗
#5 - Pain Management	2	⊗		

Figure 3.3 Physician scheduling matrix.

3.4.1. Patient Arrival Times and Rates

A patient's primary relationship is with the physician office, which will direct them to the ASF for appointments. The work flow analysis revealed two approaches by which ASF patient scheduling occurs. Approach -1: The physician schedule is divided into half-hour intervals, and patients are allotted slots on FCFS basis, with the first patient arriving 30 to 90 minutes before physician arrival. Approach-2: Using a scheduling tool the ASF projects the surgery start time for each patients and then back schedules their target arrival time. Here the authors employ a Poisson arrival process based on approach-1. The first patient for a group will arrive 45 minutes before the window start. Subsequent patients will arrive in a Poisson process with inter arrival time equal to the window length minus 60 minutes. The authors introduce the following notation:

t Scheduling blocks at the ASF, (t=1 to B)

k Physician groups active at the ASF, (1 to H)

N_k Number of physicians in group k

L_k Total number of daily patients for physicians in group k ,

E_k Number of continuous schedule blocks assigned to group k ($E_k \leq$B)

λ_k Patient arrival rate per hour for group k

Then the authors set the arrival rate such that the first patient for the group will arrive 45 minutes before the start of their first block, and the last about 1 hour before the end of their last assigned blocks. Then, λ_k is derived as follows:

$$\lambda_k = \left[\frac{L_k}{4(E_k - 1)}\right]$$

Where $L_k \leq 4E_kN_k$. Note that patient are given specific arrival times, for example 11 am, but in reality the arrival time is variant about this time. The Poisson arrival process described above integrates the inherent uncertainty in the arrival process. The early arrival is common in ASFs to minimize surgery start delay. Administrative staffing resources should be available during those appointment blocks.

Figure 3.3 illustrates the case of an ASF where H=5 groups are practicing, with $1 \leq N_k \leq 3$ for each physician group. The ASF schedule is organized into B=3 blocks with each block of 4 hour duration. As shown in Figure 3.3 the different groups have been assigned specific blocks, during which they will perform surgeries on their patients. Note that the maximum physician in any block is 6, since the example ASF has only 7 Surgery ORs. Some assumptions (i) All physicians in the group are active during the window (ii) Number of patients for each group are proportionate to their allocated capacity (iii) Surgery time and other delays may cause a physician to continue activities into the next window or into overtime, (iv) the same schedule is followed every day.

3.5. Surgery Processing Times

Central to the operating efficiency of an ASF are the processing times associated with the three key activities PreOP, Surgery OR and PostOP. Macario (2009, 2010) notes those surgical case durations are stochastic. Cases with easier-to-predict durations include ASF type standardized surgeries or specialties that operate on the body surface or extremities, such as hysterectomy, hernia repair, or cystoscopy. In contrast, difficult-to-predict cases are the more complex, nonstandard surgeries done in an in-patient setting, such as cancer surgeries or major intra-abdominal procedures. The longer the surgery, the lower the accuracy in estimating case duration. These surgeries also are more correlated to the operating behavior of a specific physician. While the authors did track some surgery times during the work flow analysis, this data is not sufficient to make reliable time estimates for the range of surgeries seen in ASFs. This research will be based on data reported by the National Center for Health Statistics. NHS Report (2009) provides national estimates of surgical and nonsurgical procedures performed on an ambulatory basis in hospitals and freestanding ambulatory surgery centers in the United States during 2006. Procedures presented are coded using the ICD–9–CM code. The ICD-9-CM (International Classification of Diseases, 9th Revision, Clinical Modification) coding system is used to code signs, symptoms, injuries, diseases, and conditions.

Surgery procedures are also coded using the CPT (Current Procedural Terminology) code. The critical relationship between an ICD-9 code and a CPT code is that the diagnosis supports the medical necessity of the procedure. Strum et al. (2003)

confirm CPT code and the associated SM code is the most important factor when predicting surgical time.

In this report Surgery OR time is defined as the time spent in the operating room during which the surgical procedure occurs. Typically, the surgical time is the time from when the process is initiated by the physician (e.g. incision) till when physician indicates process end (e.g. wound is closed). From this report 15 surgeries were selected for incorporation in our simulation model. The associated processing data for the 10 surgeries is shown in Table 3.1 which exhibits details of the surgery type with names and processing time. Note that the study reports the standard error and the authors have estimated the standard deviation using n=40. Further for all surgeries a ten minute Surgery OR capture time is added to account for the intervals before and after the actual surgery process. Commonly, this is referred to as the case duration time, which is defined as the time from "wheels in" (when the patient is brought into the room) to "wheels out" (when the patient exits the room). These none operative factors are a small fraction of the entire case duration and tend to be constant within one type of surgery.

The processing time of PreOP and PostOp activities are also shown below in Table 3.2. The NHS Report (2009) provides PostOP times aggregated for all surgeries, and does not study them as a function of the surgery code. Further no estimates of PreOp times are provided. Results from the workflow analysis were used to estimate this data. The PostOP times are recorded as Mean = 54 minutes, 25^{th} percentile = 30 minutes and 75^{th} percentile 68 minutes. Based on this data combined with the work flow analysis the PreOP and PostOP times are arranged into three lengths, which are then associated with

different surgeries in the next section. These process times include 10-15 minutes entry/exit and setup times associated with each patient transit.

Table 3.1 Processing Times for Common ASF Surgeries

Surgery #	Surgery Procedure and ICD-9-CM codes	Mean Time (Minutes)	Std. Dev. (Minutes)
1	Cataract - 366	29	4.5
2	Benign neoplasm of the colon - 2113	31	4.2
3	Diverticula of the intestine - 562	25	5.1
4	Intervertebral disc disorders - 722	32	10.8
5	Hemorrhoids - 455	27	3.2
6	Gastritis and duodenitis - 535	24	5.1
7	Chronic diseases of tonsils and adenoids - 474	31	4.8
8	Otitis media and Eustachian tube disorders - 382	21	4.6
9	Carpal tunnel syndrome - 3540	28	3.9
10	Inguinal hernia - 550	55	7.4

Source: National Health Statistics Reports Number 11 January 28, 2009–Revised

Table 3.2 Processing Times for ASF Surgery PreOP and PostOP Times

#	Pre OP Procedures	Mean Time (Minutes)	Std. Dev. (Minutes)
1	Short Preparation	20	6
2	Average Preparation	40	8
3	Long Preparation	60	11
#	Post-Operative Procedures	Mean Time (Minutes)	Std. Dev. (Minutes)
1	Short Recovery	45	6
2	Average Recovery	65	8
3	Long Recovery	90	15

Empirical studies have shown that surgery times are best modeled using a log-normal distribution. Strum, May, and Vargas (2000) conclude that lognormal distributions fit the surgery data better than normal distributions for large sets of surgery times. Consequently, the practice of estimating surgery times based on a lognormal model has been widely adopted. A log-normal distribution has positive support and positive skewedness, which is applicable to surgery times.

3.6. Patient Types & Resource Usage

The simulation literature in surgery OR modeling typically models the flow path as a function of the surgery time. Frequently, these models have considered a set of surgeries to be performed and are then exploring sequencing solutions to reduce patient wait time. The approach here is to define a set of patient flow paths that represent different combinations of physician groups, PreOP times, PostOP times, surgery codes, and the associated staffing resources usage in a typical ASF. The notation is as follows:

$\hat{\imath}$ Surgery codes performed at the facility (Surgery # in table 3.1)

i Patient types that flow through the facility

n Patient activity sequence through facility resources (1=PreOP, 2=Surgery OR, 3=PostOP)

e PreOP procedures types, $e = 1 \text{ to } 3$ (PreOP # table 3.2)

f PostOP procedures types, $f = 1 \text{ to } 3$ (PostOP # table 3.2)

g PreOP procedures length, *(1=Short, 2=Average, 3=Long)*

h PostOP procedures length, *(1=Short, 2=Average, 3=Long)*

j Staffing resource categories (1=Nurse A, 2=Nurse B, 3=Med/Tech Assistant)

$M_{j,t}$ Number of staffing resource j in block t

$\Omega_{i,\hat{\imath}}$ Patient type i has surgery code $\hat{\imath}$ (1=yes), where $\Sigma_{\hat{\imath}}\, \Omega_{i,\hat{\imath}} = 1$

$\chi_{i,j,n}$ Patient type i will utilize staff resource j during activity n (1=yes)

$\mu_{i,n}$ Patient type i mean process time during activity n

$\sigma_{i,n}$ Patient type i process time standard deviation during activity n

For the purposes of this research a set of 20 patient types was setup based on the workflow analysis data, and other data reported in the literature. Tables 3.3 and 3.4 describe the resource usage associated with each of the three PreOP and Post types. Again these are based on what was observed during the workflow analysis.

Table 3.3 PreOP Procedure Staffing Resource Usage

PreOP - e	*Staff j=1*	*Staff j=2*	*Staff j=3*
1			✓
2	✓		
3		✓	

Table 3.4 PostOP Procedure Staffing Resource Usage

PreOP - e	*Staff j=1*	*Staff j=2*	*Staff j=3*
1			✓
2	✓		
3		✓	✓

Table 3.5 describes the PreOp and PostOP lengths and types associated with each patient type, plus the associated Surgery OR resources. Each surgery code generates two patient types, with the second one representing a higher acuity or complexity level. The example ASF has H=5 physician groups, and for simplicity the authors label them as 1=V, 2=W, 3=X, 4=Y, and 5=Z. The table 3.5 data derives the resource access parameter for each patient type. That is, $\chi_{i,1,1}$ =1 if PreOP type is 2, else $\chi_{i,1,1}$ =0. Likewise the processing times are also derived. For instance, $\mu_{1,2}$ =29 minutes and $\sigma_{1,2}$ =4.5 minutes since $\Omega_{i,\hat{\imath}}$ =1 and the associated surgery time is given in Table 3.1.

Table 3.5 Patient Type Staffing & Physician Resource Usage

Patient Type - *i*	PreOP Length - *g*	PreOp Type - *e*	Surgery Procedure					PostOP Length - *h*	PostOP Type - *f*
			Surgery Code - *î*	*Staff j=1*	*Staff j=2*	*Staff j=3*	Physician Group - *k*		
1	1	1	1	✓			X		1
2	2	2	1	✓			X		1
3	2	2	2		✓	✓	Y		2
4	3	3	2		✓	✓	Y		2
5	1	1	3	✓		✓	Y		2
6	2	1	3	✓		✓	Y		2
7	2	3	4		✓		W		1
8	3	3	4		✓		W		1
9	1	1	5	✓			V		2
10	2	2	5	✓			V		2
11	1	1	6	✓		✓	Y		2
12	2	2	6	✓		✓	Y		2
13	2	2	7		✓		Z		3
14	3	1	7		✓		Z		3
15	1	2	8	✓			Z		1
16	2	2	8	✓			Z		1
17	1	1	9	✓			W		2
18	2	1	9	✓			W		2
19	2	2	10		✓	✓	V		3
20	3	3	10		✓	✓	V		3

3.7. Load Balanced Surgery Schedule

As a consequence of the wide range of surgery types and patient acuities, ASFs are challenged to develop a surgery schedule which maximizes the utilization of its staffing and facility resources while at the same time, minimizing the overtime activity and surgery overhang. The research into ASF modeling thus requires the generation of a surgery schedule. An overloaded or under loaded schedule would give skewed results, making it difficult to generalize the results across the ASF industry. For the simulation research conducted here were create a surgery for the case where H=5 and N_1=2, N_2=2,

N_3 =1, N_4 =4 and N_5 =2. The baseline patient arrival schedule design is shown in table 3.6. For the baseline problem the authors assume the staffing resources are the same for all blocks, and each patient type is associated with only one physician group. This denoted by:

$\beta_{i,k}$ Patient type *i* associated with physician group *k* then $\beta_{i,k}$ =1 else $\beta_{i,k}$ =0

α_i Total number of patient type *i* to be serviced during the day

$A_{i,t}$ Number of patient type *i* scheduled to arrive in block *t*

Note that $\sum_t A_{i,t} = \alpha_i$. The surgery load ratio for a physician group in each block is the ratio of the mean schedule surgery time and the available block capacity, this is given by:

$$\Gamma_{k,t} = \frac{\sum_i \left(A_{i,t}\mu_{i,2} \middle| \beta_{i,k} = 1\right)}{4N_k}$$

And the total patient arrivals for the group k are:

$$L_k = \sum_t \left\{ \sum_i \left(A_{i,t} \middle| \beta_{i,k} = 1\right) \right\}$$

Table 3.6 Baseline Arrival Schedule of Patients at the ASF (N=20, B=3, H=3)

Patient Type - i	Physician Group - k	Patients Scheduled/Block - $A_{i,t}$			Day Total	Group Total L_k	Arrival Rate - λ_k
		t=1 8-12 am	*t=2* 12 am - 4 pm	*t=3* 4-8 pm			
9	V	2	3	0	5	18	2.6
10	V	3	2	0	5		
19	V	2	2	0	4		
20	V	2	2	0	4		
7	W	3	3	0	6	20	2.9
8	W	3	3	0	6		
17	W	2	3	0	5		
18	W	2	1	0	3		
1	X	0	0	3	3	6	2.0
2	X	0	0	3	3		
3	Y	4	5	5	14	74	6.7
4	Y	4	4	5	13		
5	Y	4	4	4	12		
6	Y	4	4	4	12		
11	Y	4	4	3	11		
12	Y	5	4	3	12		

13	Z	0	4	4	8	26	3.7
14	Z	0	4	4	8		
15	Z	0	3	3	6		
16	Z	0	2	2	4		
Total Arrivals =		44	57	43	144		

A total of 144 patients are processed in the baseline schedule, with t=2 being the blocks with the highest load. Note that the maximum$\lambda_k = 2N_k$, and physician groups Y and Z are close to the maximum, while the others have a schedule around 70% of the maximum rate. This is typical of ASFs where one or two groups tend to dominate the schedule. The baseline staffing level is set to $M_{j,t\in B} = 6$, $M_{j,t\in B} = 5$ and $M_{j,t\in B} = 6$. Based on a 75 percentile processing time for all activities accessing these resources, plus a 15% rest time, this gives a direct resource utilization of just above 50% for each staffing resource. The facility resources are set to 10 PreOP beads, 12 Surgery ORs and 20 PostOP beds.

3.8. ASF Performance Objectives – Non Clinical

The focus of this research is on optimizing the operational (non-clinical) objectives of an ASF. The key assumption in all of the healthcare operation modeling research is that acceptable clinical performance levels are not comprised as the authors search for greater efficiencies, and that is true here also. As shown in chapter 2, simulation modeling is an active area of research in healthcare systems. The authors found that in surgery OR flow modeling the research focus is commonly on reducing patient waiting. A classical

surgery OR scheduling algorithm will consider a given set of surgery cases and then derive the sequence that will minimize wait times and maximize utilization through a set of parallel ORs. Further, these models typically consider the operating room as an integrated resource, in that all the needed staff resources are captured permanently hence do not need to be separately modeled. Additionally, there are limited constraints in physician availability. With these assumptions the systems is amenable to exact model analysis using mathematical programming techniques. Some examples include works reported by Blake et al (2002), Belien and Demeulemeester (2012), and Zhang et al (2009).

In this research the authors have opted to use a simulation approach allowing us to significantly expand the model characteristics, and bringing it closer to actual ASF practice. Based on the research the authors identify three performance objectives that are of significance in ASF analysis.

3.8.1. Staffing Costs

As noted earlier the ASF maintains three types of staffing resources, and these represent the only variable direct cost of the facility. The facility will hire a numbers of nurses and med/tech assistants all of who will be active through the daily operations. If all surgery related activities are not completed by the end of the day, then some staff will continue to work beyond the close time. This staffs are then compensated at an overtime rate. The authors introduce the following notation:

$\varphi_{j,R}$ Regular time hourly rates for staffing resource j

$\varphi_{j,o}$ Overtime hourly rates for staffing resource j

O_j Overtime hours worked by staffing resource j on a typical day

Observe that once staffing level decisions are made then the regular time staffing cost is fixed. Actual operational decisions will determine O_j for a typical ASF day.

The Association of Perioperative Registered Nurses (AORN) conducts periodic surveys of nursing salaries. Its 2011 survey showed staff nurses averaging $64,900 at the general level and $ $77,700 at the higher skill level. Based on this date the authors estimate the direct staffing costs rates as shown in table 3.7 below.

Table 3.7 Estimated Hourly Staffing Resource Costs Rates

STAFF RESOURCE CATEGORY *(j)*	REGULAR RATE $\varphi_{j,R}$	OVERTIME RATE $\varphi_{j,O}$
(j=1) Nurse Group - A	$ 28	$ 40
(j=2) Nurse Group - B	$ 21	$ 31
(j=3) Med/Tech Assistant	$ 17	$ 25

After some surveys from ASFs, staffing members should get paid extra 50% (average level) more than the regular salary rate if they work after regular time.

3.8.2. Patient Waiting Time Costs

Patient wait time is a widely studied objective in many healthcare systems engineering research projects. The basic premise is that patients would want to wait a minimum time, and are inconvenienced when the wait becomes progressively longer. Healthcare is a service industry in which patients flow through a series of healthcare processes, as a result patient waiting is inherent in the system. The literature identifies two types of

patient waiting time (Gupta and Denton 2008, Liu et al 2010): (i) Indirect waiting - times between the day patients call to schedule a surgery or appointment and the actual appointment date, (ii) Direct Waiting – Scheduled start of surgery or appointment and actual start. In this research the authors model only the direct waiting time. Papers that deal with direct waiting time typically consider it along with other objectives by minimizing an objective function that is a weighted sum of a subset of these various performance measures.

Krueger (2009) estimates that Americans age 15 and older collectively spent 847 million hours waiting for medical services to be provided. He notes that patient waiting time is an important input in the health care system. Failing to take account of patient time leads us to exaggerate the productivity of the health care sector, and to understate the cost of health care. Laganga and Lawrence (2007) not that healthcare facilities frequently overbook their capacity a common cause for increased patient waiting times.

Clearly, the penalty for a patient's waiting time is another virtual cost since ASFs don't really pay for it. This is also referred to as a welfare cost. However, ASFs operate in a highly competitive market and are looking to improve their service efficiency through less waiting time. While patient waiting time is used in a wide variety of healthcare analysis models, there is little data on what the cost rate is and what its functional nature is. From the review the authors summarize that there are three possible approaches to characterize the time function nature of the patient waiting time cost curve:

(i) *Linear Waiting Cost:* A direct product of the waiting time and a waiting penalty

(ii) *Squared Waiting Cost:* A weighted square of the waiting time and a waiting penalty

(iii) *Step Function Waiting Cost:* Described by an increasing staggered step in fixed cycle

The authors introduce the following notation associated with the linear waiting cost model:

ϕ_P Patient waiting time penalty rate - \$/hour

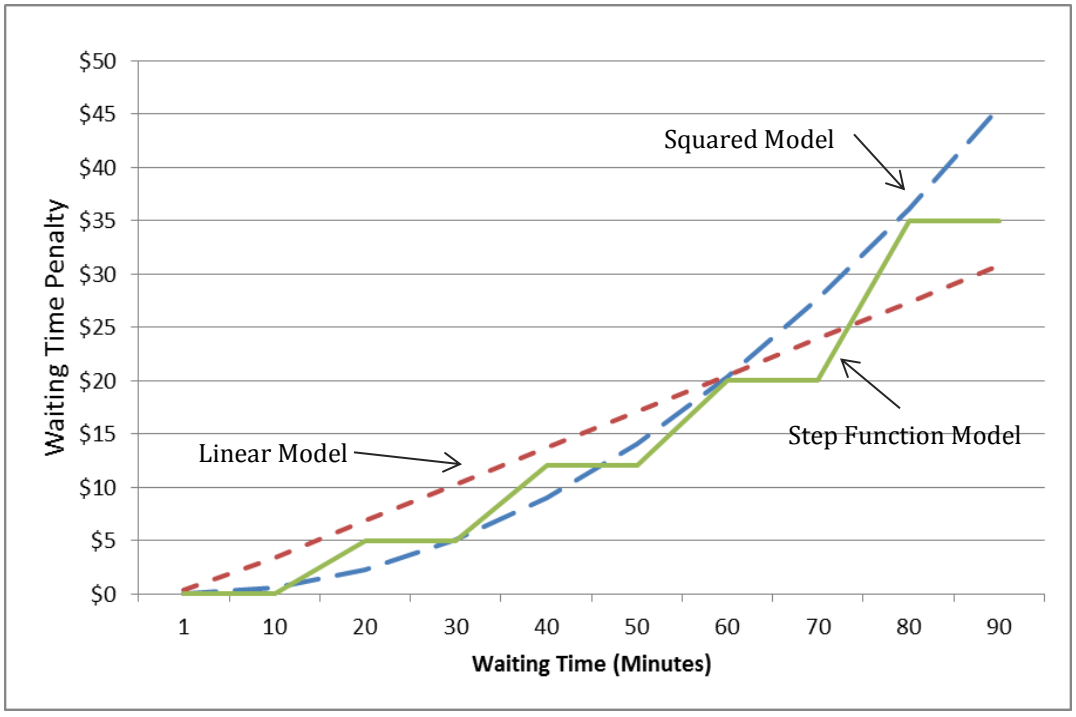


Figure 3.3 Patient time cost models.

Figure 3.3 illustrates the three cost models for the case where ϕ_P =\$20. The three models are then set such that they are benchmarked to the congruent cost of \$20 at the 1 hour time point. Observe that in the sub 1-hour the linear model emphasizes patient waiting time, while the squared model emphasizes the cost in the plus 1-hour range. The

literature review indicates that the linear model is more widely used in research analysis, while in the healthcare economics literature there is a preference for the other two models.

The authors also observe, though, that there is little discussion in the literature on the actual value of ϕ_P. A common approach is to assume the ratio $\phi_P/\beta_{j,R}$ instead. More frequently the authors see anecdotal mention of ϕ_P in news articles. Most of these recommend using some standard labor rate as a surrogate for ϕ_P. Princeton economics professor Alan Krueger argues in a widely cited NY Times article that ϕ_P should be set equal to U.S. average hourly wage of the private nonsupervisory non-farm payroll ($20.25 for 2013). Agarwal (2012) at the University of Maryland's Center for Health Information and Decision Systems goes further and say's it should be equal to the average wage of the entire non-farm payroll ($24.08 for 2013). There is also a school of thought that wages already account for waiting times in that workers are eligible of sick days etc., and should therefore not be a cost since there no real wage lost. The conclusion is that the best cost model for ASFs is a linear model with a cost rate discounted from average non-farm payroll. The authors thus set ϕ_P = $17.50 a 25% discount from the average of $24.05. A key motivation for this is that surgical settings even short waits are uncomfortable for the patients.

3.8.3. Physician Delay Costs

Physicians are the most valuable and critical resource in any healthcare facility. Quoting from Agarwal (2012) “A physician's time is perhaps the scarcest resource in our health care system and needs to be utilized optimally. Doctors play a noble role in our society — they save lives and relieve pain. Their time is valuable. And it is preferable if the patient waits rather than the doctor”. In our research on the current operational flow of ASFs the authors found that the primary operational concern of ASF managers was physician satisfaction. While quality of resources and facilities are key components of physician satisfaction, timely completion of all surgeries is of primary concern. Quantifying this time cost though can be challenging. The authors start with the appointment scheduling literature. In most patient appointment scheduling models the objective is to minimize the weighted sum of three costs: patient waiting cost, doctor’s idle time and overtime costs (Zacharias and Pinedo, 2013). In the literature most papers avoid an explicit mention of the physician cost and rather develop their model to use a cost ratio defined as the ratio between the patient waiting cost and physician idle time (Robinson and Chen, 2010). Frequently, this ratio is set in the 10-20% range.

The investigation reveals that in ASFs the physician idle time is not the metric of focus, but rather the physician delay. As noted in section 3.4 physicians contract for and are assigned a set of surgery blocks. The physician group then schedules a set of surgeries to perform in their allocated blocks such that $\Gamma_{k,t}$ is less than a contract maximum, for example $\Gamma_{k,t}$ <0.65. The assumption here is that in an efficient ASF the group can service all patients in the block. Figure 3.4 illustrates the actual flow of operations. Patients are delayed in the processes leading up to the surgery starts. Additionally the surgery is delayed due to either lack of resources and/or the Surgery OR being unavailable. As a

result at the end of the block group k has not completed all surgeries and the physician has to continue working past the end time. This extended duration is the physician delay. The physician considers this delay to be the responsibility of the ASF. The authors find that surgery physicians are very sensitive to this delay since it has a tandem effect on their sequential activities. Dissatisfaction with this delay could cause one or more doctors in a physician group to take their patients to a competitive facility. Physician delay is therefore a business opportunity cost that an ASF must consider in planning its operations.

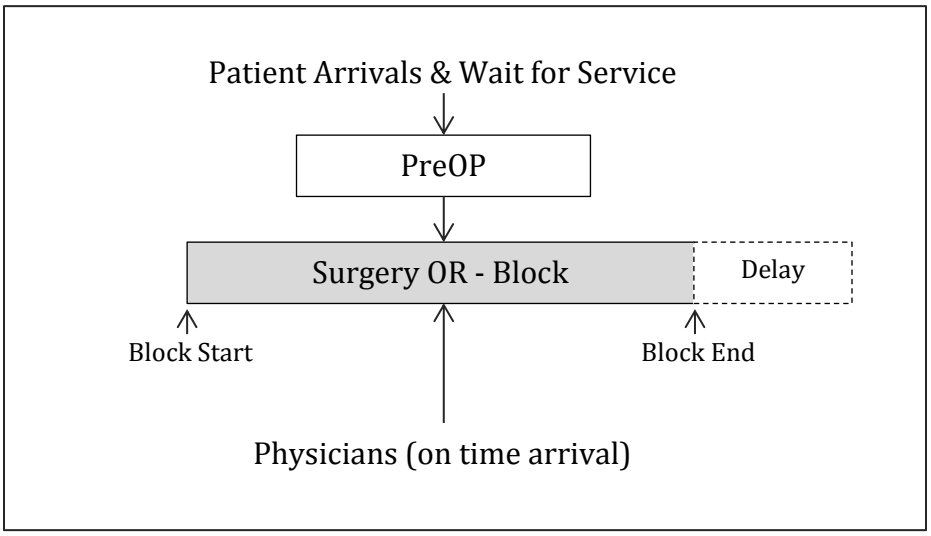


Figure 3.4 Physician delay explanation.

The authors assume here without loss of generality the physician delay penalty rate is the same for all physicians in all groups. Introducing:

ϕ_D Physician surgery block completion delay penalty - $/hour

To estimate the value of the authors start with the average hourly earnings for a general surgeon, a practice most representatives of ASF physicians. The Medical Group Management Association's physician compensation and production survey (MGMA Report 2011) estimates annual compensation at $265,000 or $155/hour, assuming a workload of 1700 hours/years. Applying a penalty factor of two for the delay impact here the authors set ϕ_D =$300 for the ASF analysis.

The ASF operational objective function can then be derived as the daily sum of three costs (i) Staffing – both regular and overtime costs (ii) Patient waiting time costs and (iii) Physician delay costs. Using the notation introduced above the cost objective then is:

$$Minimize: \quad \Omega = \sum_j \sum_t \left(4\varphi_{j,R} M_{j,t}\right) + \sum_j \left(\varphi_{j,O} O_j\right) + \phi_P T_P + \phi_D T_D$$

Where,

T_P Total waiting time all patients entering ASF in a day

T_D Total delay time for all physicians active at the ASF in a day

The first term in the objective is deterministic, that is once a decision is made on the staffing levels ($M_{j,t}$) then this cost is directly calculated. The other three costs are stochastic in nature since they are dependent on three system performance variables - O_j, T_P and T_D. A key research question is how to derive an accurate estimate for these variables.

3.9. Deriving O_j, T_P and T_D

The key variables in deriving the cost objective Ω are the variables $M_{j,t}$, O_j, T_P and T_D. Of these $M_{j,t}$ is a management decision and becomes a fixed cost once decided. The other three are operational outcomes and have to be derived either analytically or by the use of a simulation model. Key factors which limit the application of mathematical programming methods in healthcare setting include (i) flexible and complex flow paths, (ii) multiple classes of patient entities (iii) multiple floating and fixed resources (iv) uncertain services times and (v) scheduled but uncertain patient arrivals. (Marcon and Dexter (2006)) state that dynamic simulation is one of the best ways for studying the performances of healthcare systems. Denton et al (2006) observe that while the single OR scheduling problem can be optimized using stochastic linear program, multi OR problems are much more complex and can only be analyzed using simulation model. In a comprehensive overview of the outpatient appointment scheduling literature (Cayirli and Veral, 2003; Westeneng, 2007) the authors see that of 23 papers, 17 apply a simulation method to solve the problem. The approach here is also to use a discrete event simulation model to derive accurate estimates of O_j, T_P and T_D.

3.10. ASF Simulation Model

The simulation model was built and implemented on the ARENA 14 platform. ARENA is a well-known and popular discrete event simulation software platform. ARENA uses a graphical interface allowing the user to build model by placing functional modules that represent pre-coded processes or logic in a flow system. Connector lines are used to join

these modules together and specify the flow of entities. While modules have specific actions relative to entities, flow, and timing, the precise representation of each module and entity relative to real-life objects is subject to the modeler. Statistical data, such as cycle time and WIP (work in process) levels, can be recorded and outputted as reports. ARENA has been used by many healthcare process analysis research groups.

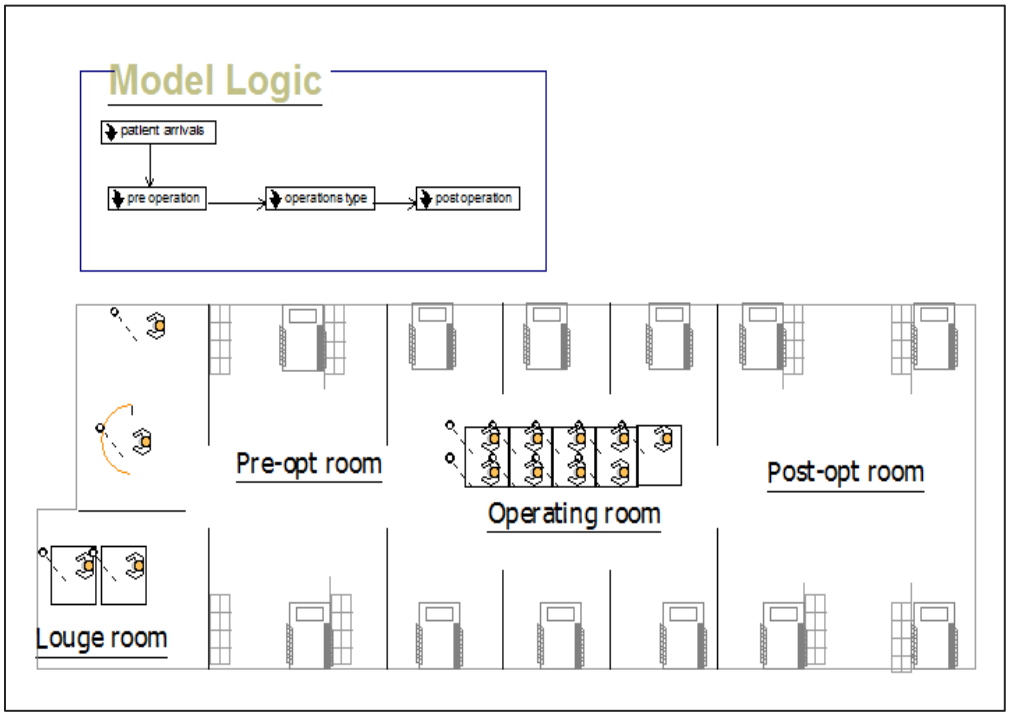


Figure 3.5 an ARENA simulation animation layout for an ambulatory surgical facility.

The model was developed in the windows platform and all experiments were conducted in this platform. Figure 3.5 illustrates the visual interface of the program which provides an animation screen of the ASF operations. The animation mode can be used to help users better understand the ASF models operations, particularly when complex patient flows are involved. The animation also helps in program debugging.

3.10.1 – Patient Arrivals through Registration

The patient arrivals activities include (i) patient tagging by type of physician group and (ii) registration process. The flowchart describing the overall process is shown in Figure 3.6. Physically these activities occur in the ASF lounge. The lounge is capacity unbounded and can accommodate all arriving patients. Two arrival processes are followed:

(i) A Poisson arrival process generated within the ARENA model in which case patient arrivals are independent both by physician group and patient type. The process is controlled the logic and parameters described earlier, which is characterized by patient type arrival independence.

(ii) An externally generated fixed arrival schedule that is entered through an Excel file identifying patient type and arrival time. In this case arrivals may or may not be independent in terms of physician group or patient type. It will depend on the rule by which the arrival sequence is created.

The common process through which all patients will go is registration following which they enter the ASF. Registration is modelled as a basic M/M/1 queue with dedicated resources, that is they are captive to the server.

3.10.2 – Pre Operation Process

Includes the activities of (i) assigning PreOp Type, (ii) Queue and capture of PreOp bed resource (iii) Queue and capture of needed staffing resources (iv) Execute PreOp process and (v) Block PreOp bed resource while in queue for surgery bed resource. Since the PreOP bed resource is capacitated two logic blocks are created, one to manage the PreOp

queue and the other to generate the PreOp parameters once the bed resource becomes available (Figure 3.7). The PreOP activity block will delay process start unit the corresponding nursing resources are captured. Figure 3.7 shows the logic sequence for a specific PreOp type. Those blocks as explained stand for different sub processes which could have different parameter settings covering process time and related distribution, resource needed and queuing related types and those lines connect between the ins and outs. At the end of the process time the nursing resources are released, but the PreOp bed resource remains blocked till a Surgery OR is available, at which point the patient will enter the surgery process.

3.10.3 – Surgery Process

Includes the activities of (i) assigning Surgery Type, (ii) Queue and capture of Surgery OR resource (iii) Queue and capture of needed staffing resources (iii) Queue and capture of associated physician resource (iv) Execute surgery process and (v) Block Surgery OR resource while in queue for PACU bed resource. The flowchart describing the overall process is shown in Figure 3.8. Similar to the PreOP bed resource the Surgery OR resource is also capacitated and modelled likewise. The Surgery activity is setup as a Seize-Delay-Release block which will delay process start unit the corresponding nursing resources and physician resources are captured. Figure 3.8 shows the logic sequence for a specific surgery type. The Surgery OR also includes a fixed time clean-up process between surgeries, which is embedded in the logic in the end of the process time, the nursing resources and physician resource are released, but the Surgery bed resource remains blocked till a PACU bed resource is available, at which point the patient will enter the PACU process.

3.10.4 – Post Anesthesia Care Unit (PACU) Process

Includes the activities of (i) assigning PACU Type, (ii) Queue and capture of PACU bed resource (iii) Queue and capture of needed staffing resources (iii) Execute PACU process and (v) Release patient from ASF. The flowchart describing the overall process is shown in Figure 3.9. A key difference between PACU and PreOP or Surgery is that nursing resources are not captured for the entire process. Rather, they are used for an initial setup period and a final release period. Thus two queue and capture processes are needed. Similar to the Surgery OR the PACU activity is setup as a Seize-Delay-Release block which will delay process start unit the corresponding nursing resources are captured.

3.10.5 – Staffing and Physician Resource Control

Both physician and staffing resources are modelled as floating resources. That is they are not captive to any specific server or activity block. Capture times also vary by PreOp type, Surgery type and PACU type. As shown in Figure 3.10 logic blocks are programmed to link the various ASF activities to the resources. Delays are also setup to control the flow of staff resources between activities.

The Figure 3.11 is in the “Statistical” module where you can create special output file. In the “expression builder” where you can pick up existing value codes or combine those codes to be your simulation output value just like the option box on the left in the chart. In addition, you can label the file for the special output and save it in the wanted place by checking the left two columns. “Building the model”→”verify it”→”error found”→”modify” is a system loop until the model reaches the final requirement. All the examples of easy set ups cannot be displayed here all and its original version is the

example model called “emergency room” from Arena official install package. However, because of different logic behind the original and current, they are totally different two models except the similar animation layout.

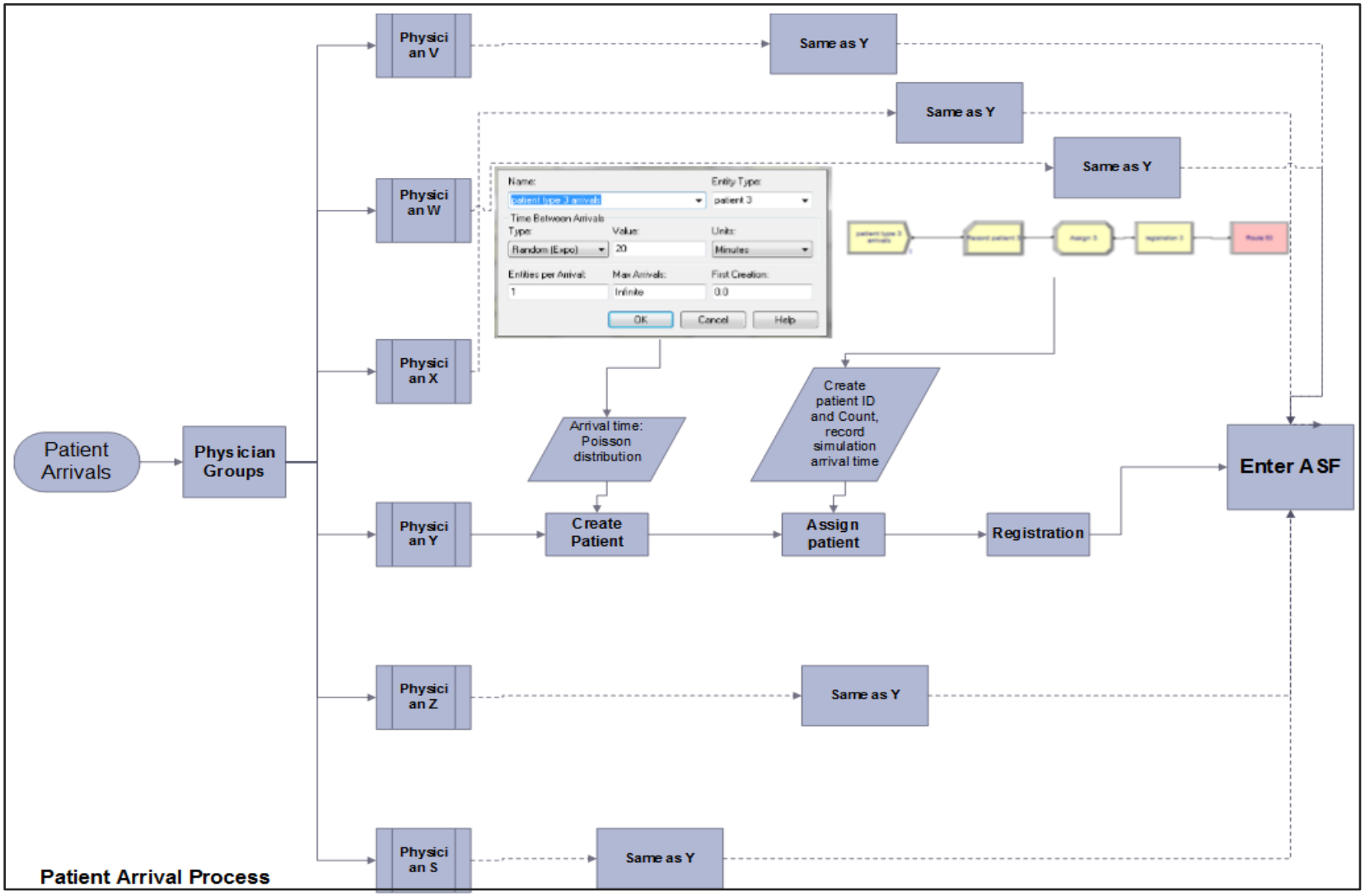


Figure 3.6 Model flowchart patients arrivals to patient assignment.

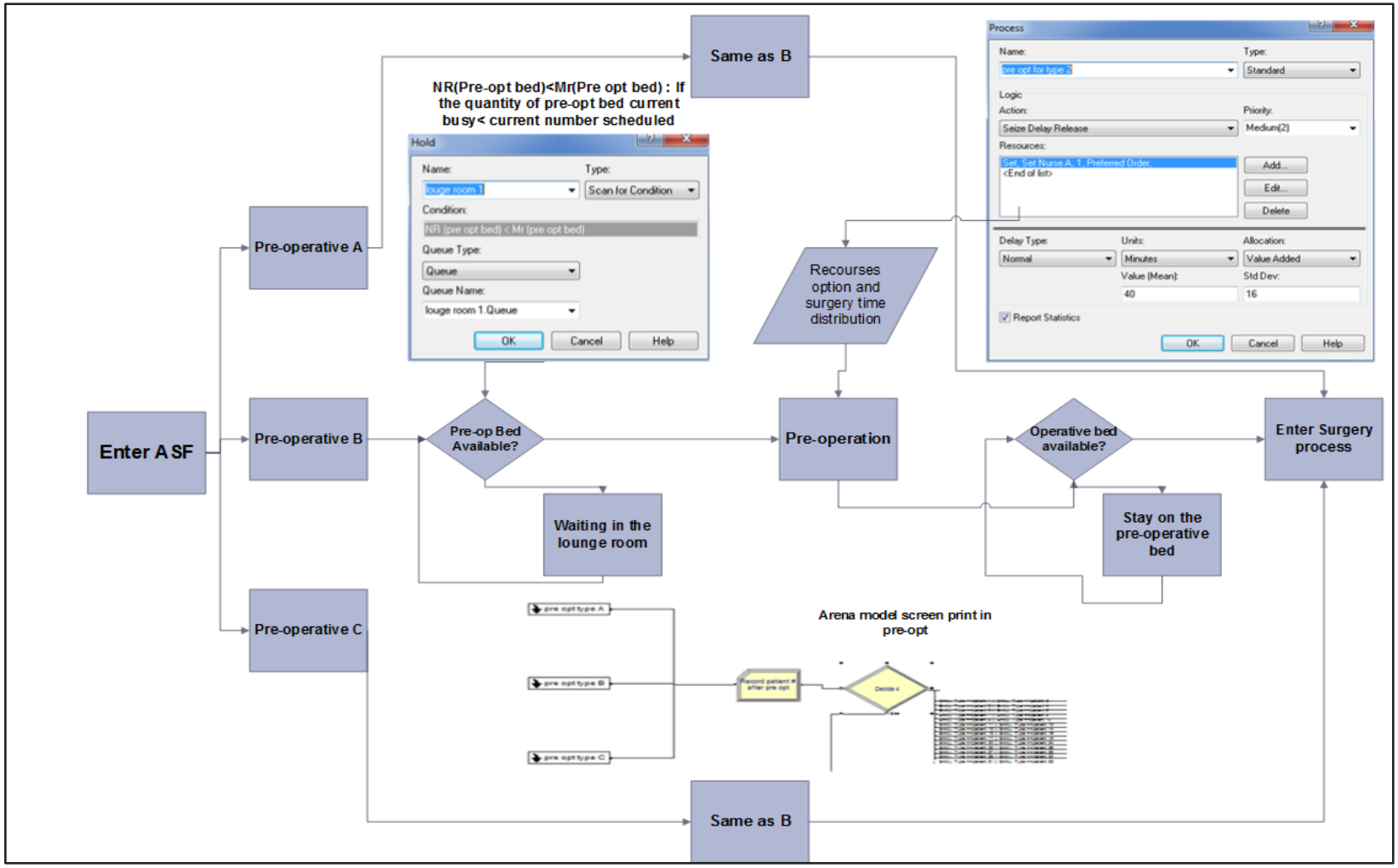


Figure 3.7 Model flowchart patient assignments through pre operation.

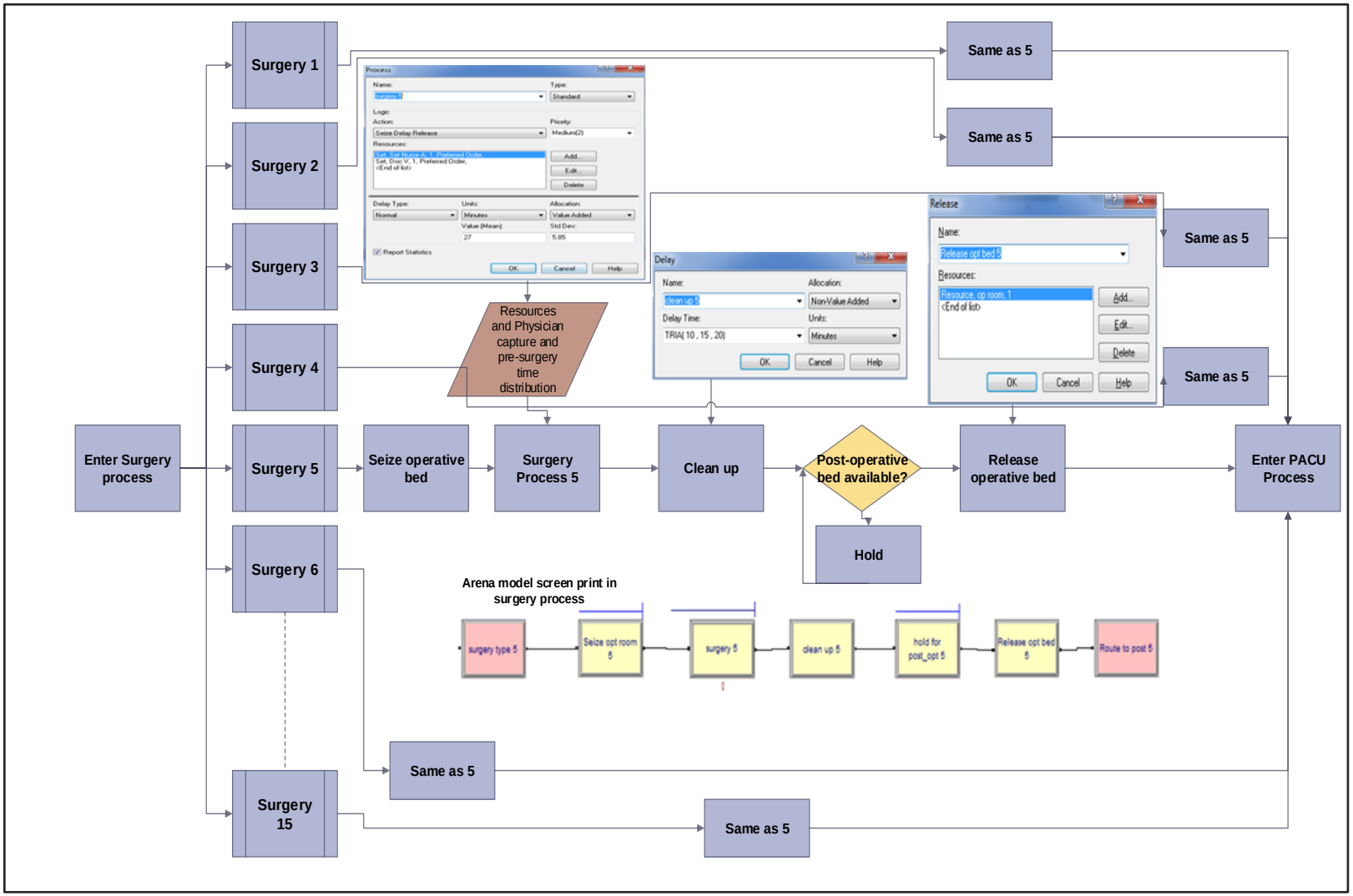


Figure 3.8 Model flowchart surgery activity in OR.

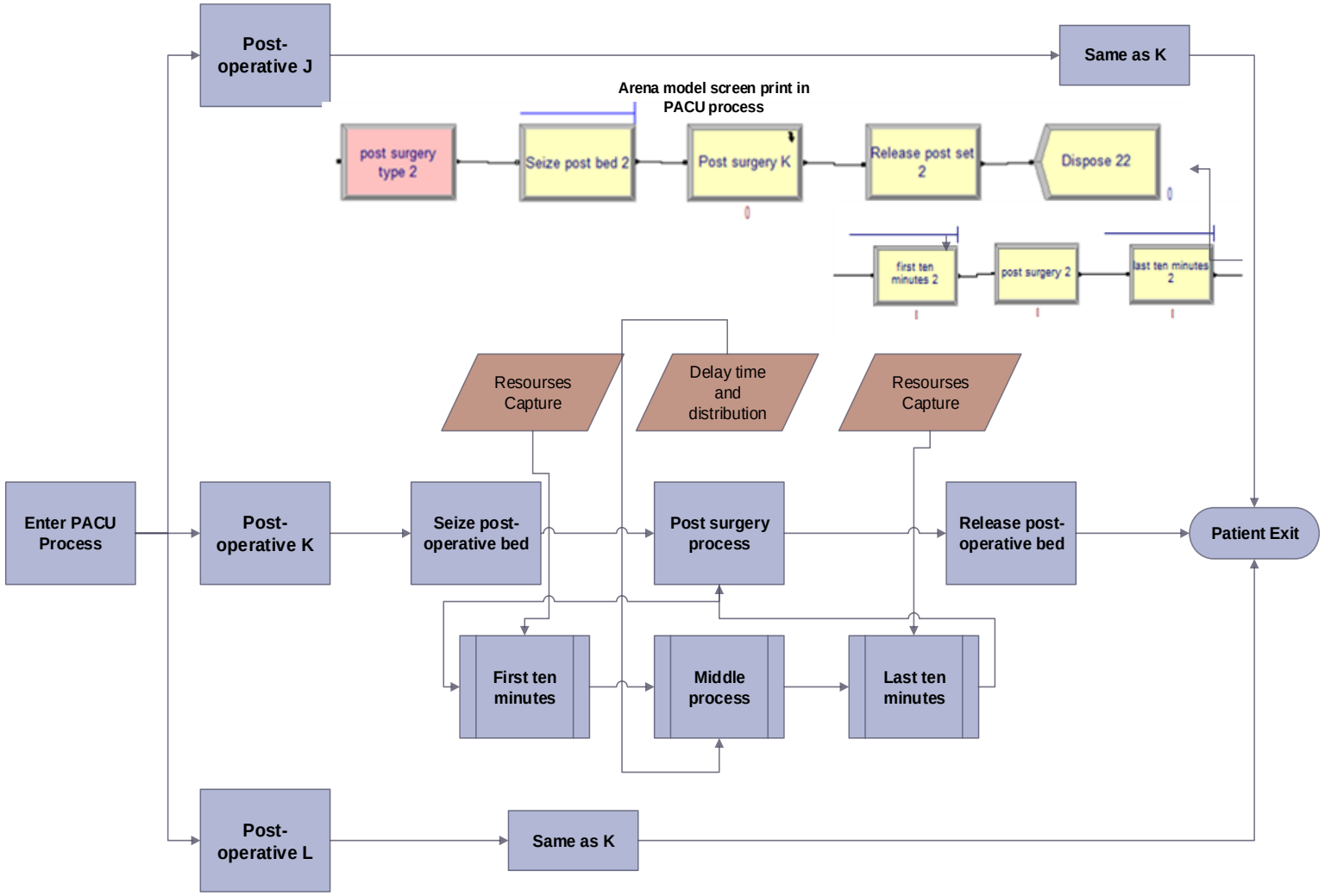


Figure 3.9 Model flowchart post operation activity in PACU.

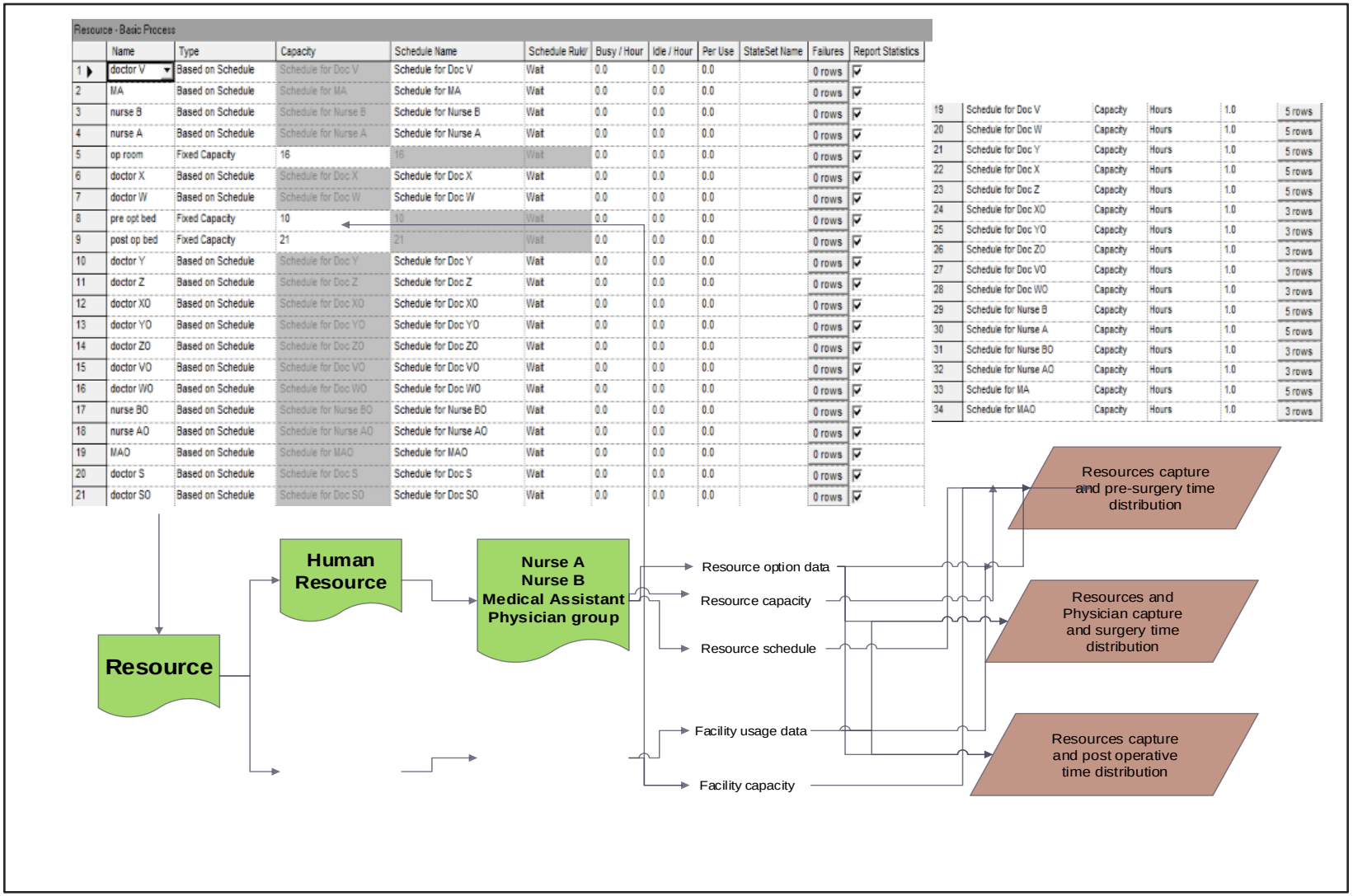


Figure 3.10 Model Flowchart for Floating Resource Allocation.

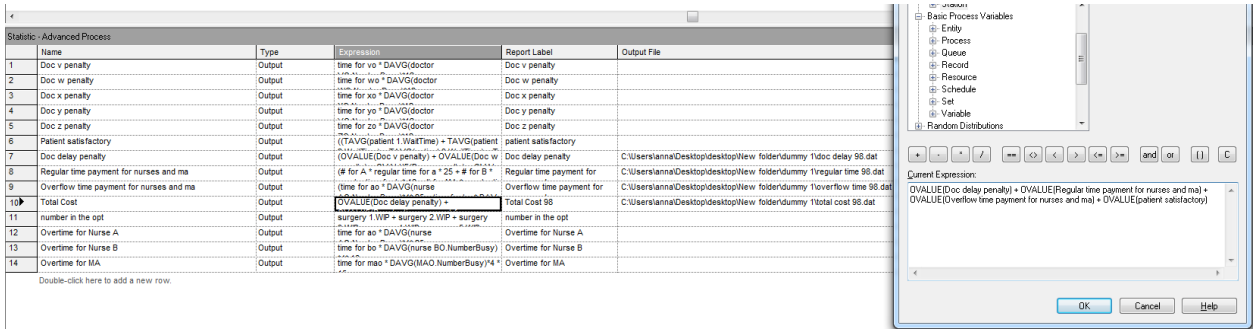


Figure 3.11 Arena statistical module.

3.11. Model Process Validation

To have valid experimental data is an important task for further results analysis; therefore several statistical methods have been applied to confirm its logic correction. Some attached functions from Arena can help us track the real numbers in process when the authors change inputs. Firstly a real time clock (on the upper left corner of Figure 3.12) has been added to the system to check its working time and relate to those numbers accordingly. The numbers the authors picked up here basically from two main aspects: patient and staffing members (the number of doctors and facilities are fixed). The Work in Process (WIP) number of patients has been tracked to ensure the completion of all arranged surgeries per day. Because one of the optimization focus may be concerning on best staffing levels, the three types of staffing members are also listed here with two statuses (busy and idle) and total number, from which you can track what will be the enough level for staffing. Several extreme cases have been studied (listed below) and all the results have proved the validation of the data.

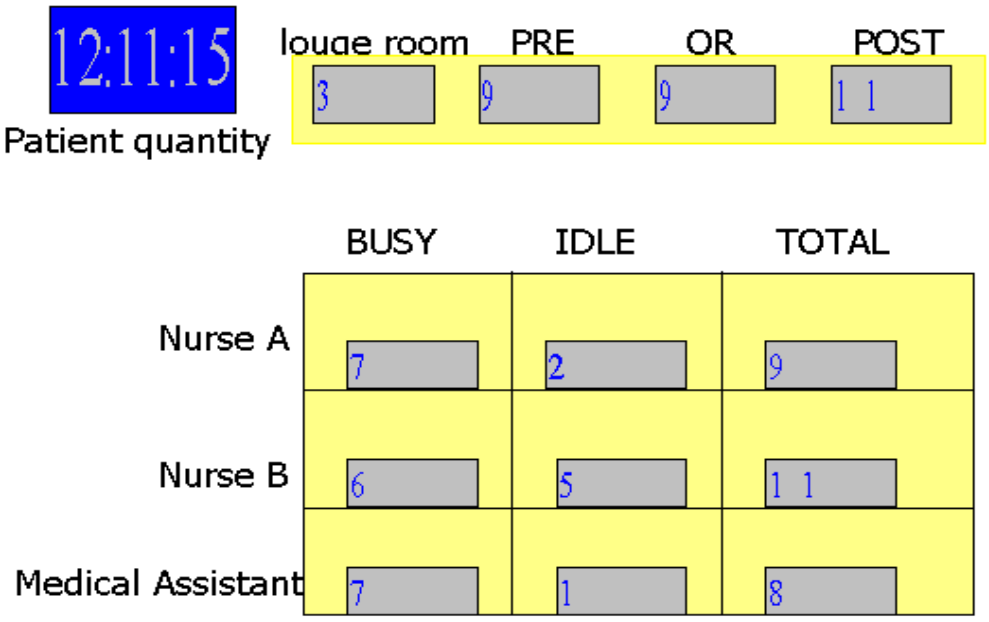


Figure 3.12 Arena animation.

- Patients' quantity, path through the process:

The total number of patients which have gone through the simulation model will be recorded by the software and shown in the results report. By manually calculation, you have some planned number once you input the data, see if the total number matches the results quantity, there may be some patients still in some process after the end simulation time, so the quantity you planned for the model should equal to the WIP patients add final out patients.

To track the paths of patients is another validation of model, and the software itself gives us shortcuts to do it by observing the dynamic running process. Since the patients are in groups belonging to some physician group, specify one group and mark them as different animation icons, from the planned path combined with the time, you

could get the paths for the patients in ASF, by waiting in some preoperative, operatives or postoperative and checking with special icons you put and even the time and the number of them gives you an immediate confidence all the paths setting up correctly.

- Resource quantity and utilization

As shown in the chart 3.6, total resources quantity have been tracked during the simulation process could immediately give you an idea that the number you put is more or less than needed. However, not from running the simulation, the expected number of enough resources could be calculated through the input data by finding out the number and the time of patients who are assigned to need that resource. But the number calculated is just the estimation under certain variance and confidence level so usually the calculated number will be initially put in the model and by running and adjusting several times the final results could be confirmed to be corresponding to the planed input quantity. For the utilization's estimation is just combined with the quantity has been put in the system, however, since all resources have an effects on each other, hardly the authors could expect how much the utilization could drop from adding more resources, but the trend of utilization should be composite with their quantities.

- General extreme cases

Zero staffing members: since staffing members are in need for different process here and there, the expectation result for the number of total patients out should be 0. By setting members to be zero separately, only part of the patients can go through upon simple reasoning consequence. Zero physician members in groups: five physician groups are planned to operate ten types of surgeries for twenty types of patients, thus separate

zeros in each group would lead to lacking of patients in that group accordingly and other patients won't get affected supposedly. But when all doctors are disappearing from the system, no patients will be helped but trapped waiting for the doctors in the operating room.

Zero facilities: three types of beds are set to be zeros individually first, since all beds are necessary for any process except registration, all patients are supposed to be stuck at matching process which lacks of beds and the queue for the beds is accumulating fast.

Above three extreme validation methods are basic levels' strategies and all of which are stopped by Arena with the warning that too many entities (patients) in one module which exceeds the original setting. Other factors like processing time in mean and variance could also be changed to check the validation of the results by using almost the same ways: compare the results from extreme cases with the expectation conclusions. In the meantime, lots of variables you can setup or pick up from the software package, and the current simulation statuses are easy to be read from the numbers, graphs or even some elaborate charts after times of modifying. For some tracked records are about patients' arrival time in between, resulting in blur in distribution identifications, Arena also has its data analysis function helping organizing input data and finding proper distributions. Additionally, it has automatic breakpoint to pause at any conditions you set up.

To conclude, either by manually calculating the expected number and then comparing with the simulation results, or by using help tools attached from Arena

software package, lots of work has been done behind until the model finally could be setup for running experiments.

3.12. Potential Decision Making Problems

The purpose of this research is to develop decision making models that allow us to optimize the performance of ASFs. The authors have formulated the operational structure of the problem and described a new objective function which accurately represents ASF practice. Our analysis reveals several ASF problems than can effectively and efficiently be solved using the model developed here. The authors introduce them here:

Optimizing Staffing Resources Levels – As noted earlier the variable largest direct cost in an ASF is the staffing cost. In our interaction with ASF facilities this was a key management concern. Current practice, involves manual expertise whereby a person with staffing experience will make decisions on staff levels typically for the upcoming week. ASG operators need decision models that can characterize the relationship between staffing levels and operating costs, and consequently prescribe optimal staffing levels.

Assignment of Schedule Blocks to Physician Groups – In section 3.xx the authors introduced the block scheduling arrangement that ASFs negotiate with physician groups. Since many schedule combinations are possible, ASF need models that can predict the performance impact of the combinations. Further, they are looking for assignment rules to derive block schedules which optimize performance.

Specifying Patient Arrival Schedules – The common approach to patient scheduling is to setup a uniform arrival pattern in which patients are scheduled to arrival at a constant mean rate. Frequently, then mean inter arrival time is 45 minutes which is then factored as described. ASFs are keen to learn of alternative scheduling method which dynamically adjusts the arrival rate so as to optimize performance. This problem is of much interest in physician office visit scheduling, and the literature is rich with many models. ASFs need models which address this problem specific to their operating structure.

Patient Arrival Sequences – This problem is an extension of the schedule problem. Typically, patient schedules are generated without considering the specific type of surgery to be performed. If a physician group performs only type of surgery then the sequencing problem is mute. But when multiple surgery types are performed then the OR scheduling literature shows that classical machine sequencing rules can be used to improve performance. ASFs need sequencing rules that can be used in conjunction with patients scheduling methods.

Physical Resource Capacity Levels – Another major cost of the ASF are the physical resources, specifically the ORs, PACU beds and PreOP beds. While these are fixed capital costs, the ASF does have the option of activating and deactivating these resources as needed. These actions will involve some kind of setup cost and possibly a maintenance cost. ASFs need models which can prescribe a strategy for managing these resources.

CHAPTER 4

ANALYSIS OF THE STAFFING STRATEGY IN AN ASF FACILITY

In chapter 3 the authors introduced a new objective function for evaluating the operating performance of ASFs. A simulation model to track this objective in an ASF was also developed and presented. In this chapter the research transitions to investigation of decisions and solutions which can be utilized to improve the operating performance of an ASF. In this chapter the specific focus is on *Optimizing Staffing Resources Levels.* As noted earlier the variable largest direct cost in an ASF is the staffing cost. In the interaction with ASF facilities this was a key management concern. Current practice, involves manual expertise whereby a person with staffing experience will make decisions on staff levels typically for the upcoming week. ASF operators need decision models that can characterize the relationship between staffing levels and operating costs, and consequently prescribe optimal staffing levels. *Question:* What is the staffing level for Nurse-A, Nurse-B and Medical Assistant that minimizes the ASF Performance Goal? *Solution:* Use a simulation experimental approach to determine the staffing level. Decision variables are:

$M_{j,t}$ Number of resource j in block t

This chapter is organized as follows. Section 3.1 defines all the resources in the model which will be used in later model constructions; The second section (3.2) of this chapter is about model constructions in details to a general ASF including: ASF operating process analysis with assumptions and an event flow chart clarify logic connections

between process; (3.3) introduces model input data in different tables under one scenario and general performance objective function set up is in (3.4); (3.5) states that rational reason of choosing discrete-event simulation(3.5.1) and key surgical processes converting into Arena model (3.5.2); statistical validation of the simulation model is in (3.6); and the last section of the chapter (3.7) concluded causes of uncertainties in ASF system and those changeable decisions which the authors could make to optimize ASFs. The listed topics in the conclusion will be analyzed in details in following chapters.

4.1. Defining the Staffing Problem

An ASF maintains three staffing resources (j = 1 to 3) which together account for the primary direct cost of the facility. Clearly then the ASF attempts to reduce the staffing levels. Variations in staffing levels though are inversely related to two other objectives patient waiting times and physician delays. The ASF staffing problem can then be defined as follows:

$$Minimize: \quad \Omega = \sum_{j}\sum_{t}\left(4\varphi_{j,R}M_{j,t}\right) + \sum_{j}\left(\varphi_{j,O}O_{j}\right) + \phi_{P}T_{P} + \phi_{D}T_{D}$$

The decision space is: $1 \leq M_{j,t} \leq M_{j,MAX}$ for j = 1 to 3 and t = 1 to 3. Where $M_{j,MAX}$ is the maximum assignable staffing resources and $M_{j,t}$ is integer. As noted earlier in section 3.9 Of these $M_{j,t}$ is a management decision and becomes a fixed cost once decided. The other three are operational outcomes and have to be derived either analytically or by the use of a simulation model. Clearly there is an inverse relationship between the effect of $M_{j,t}$ and the operational outcomes. Thus the authors would expect a

convex relationship between $M_{j,t}$ and Ω. In this chapter the authors investigate this relationship and used it to prescribe optimal values of $M_{j,t}$. The research strategy is to create a series of $M_{j,t}$ decision scenarios to track Ω for a specific ASF example. This is repeated for additional problems to detect a generalized trend.

4.2. Experimental Strategy to Determine $M_{j,t}$

Simulation models allow decision optimization under stochastic conditions. For simple problems, analytical techniques can be applied (Ross, 2003). Those analytical techniques become inapplicable when the problem gets more complicated. In these cases a simulation based optimization approach has been shown to be a powerful tool (Kao & Chen, 2006). Modern simulation platforms typically include a black-box parameter optimization tool. ARENA integrates an optimization toolbox OptQuest (Glover et al., 1999) which contains several scenario and configuration analysis algorithms (mainly meta heuristics). Figure 4.1 provides a classification of the various possible optimization approaches.

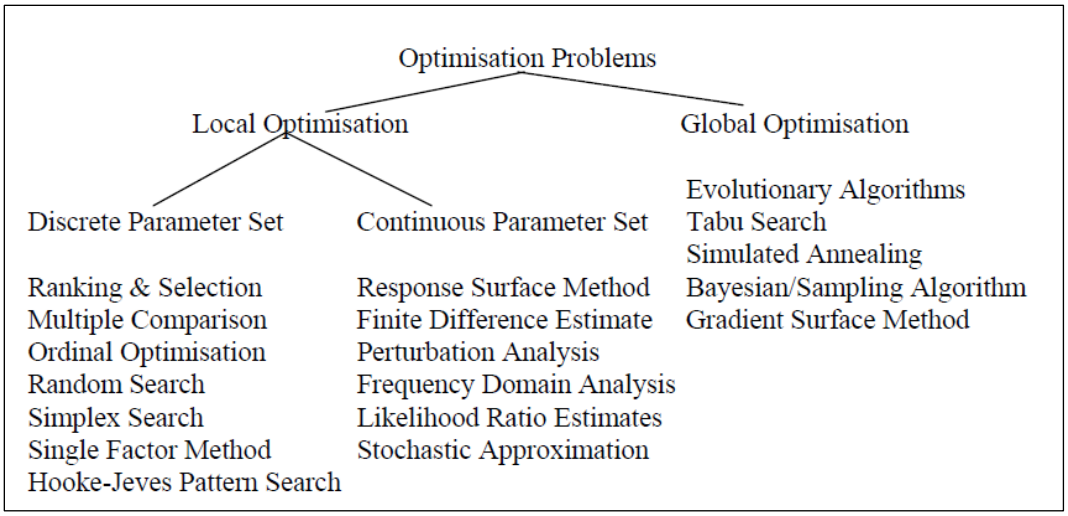


Figure 4.1 Classification of grouping of simulation optimization approaches.

Source: (Tekin & Sabuncuoglu, 2004)

4.3 Design of Experiments

Given that multiple resource types are involved the approach is to design a multi-factor experiment to derive $M_{j,t}$. The developed ASF simulation model uses classical scenario analysis to capture the relationship between the decision factors and the performance objective described above. Each scenario is represented by a simulation experiment. Each scenario is represented by a unique combination of staffing levels. In general usage, DOE or experimental design is the design of any information-gathering exercises where variation is present, whether under the full control of the experimenter or not. Unlike the one factor test which changes one factor at a time while keeps others constant, DOE provides a full insight of the interaction between design elements rather than individual effects. To develop the experimental strategy the authors introduce the baseline problem as introduced in chapter 3 with B=3, H=5 and P=20. Key data for the baseline problem

are introduced in Figure 3.3 and Tables 3.1 to 3.7. Additionally, the performance function cost coefficients are ϕ_P = \$24.05 and set ϕ_D = \$300.

4.3.1. Selecting the Experimental Array/Space

To determine the expected staffing resource utility and other performance parameters resulting from staffing levels, the authors conducted a full factorial simulation study. As noted above at the starting point there are 9 decision factors for the baseline problem: $1 \leq M_{j,t} \leq M_{j,MAX}$ for j = 1 to 3 and t = 1 to 3. Several initial experiments were conducted on the baseline problem. Based on the observed sensitivity of the performance measure to the factors it was decided to trim the experimental space. Specifically, (i) The staffing level was the same for all time periods, that is $M_j = M_{j,t}$ for t = 1 to 3 and (ii) the staffing level for medical assistants was predetermined at M_3 = 10 and therefore not an experimental factor. For the remaining two factors the authors set $M_{1,MAX}$ = 9 and $M_{2,MAX}$ = 11. Beyond these levels the authors see sharp increases in Ω. The experimental array is then shown in Table 4.1 Later the authors will add two new problems to validate the conclusions, and for these problems the experimental array is similarly derived. Note that the decision space is discrete.

Table 4.1 DOE Experimental Array for Baseline Staffing Problem

Expt #	1 to 5	6 to 10	11 to 15	16 to 20	21 to 25
M_1	5	6	7	8	9
M_2	7,8,9,10,11	7,8,9,10,11	7,8,9,10,11	7,8,9,10,11	7,8,9,10,11

Patient wait times W, doctor delay, and staffing overtime were the dependent variables measured during the simulations.

4.3.2. Replication Estimate for the Experiments

Simulation experiments are inherently characterized by errors or measure variance. For a valid study the simulation replication number should be estimated to get more accurate experimental results. Half width is reported by Arena as a term for variance, and "acceptable variance" is defined to be < 4% of the changeable value. If initially 100 replication number has been set and the reported half width is 175, while 4% of changeable value (which is the total value subtract the fixed regular salary payment) is 45, then the required replication times should be far more than 100. To estimate the valid number of replications under certain half width, the following definitions and equations are used. Standing as the most direct output value, half width is just showing everywhere after mean value in the simulation reports. If a value is returned in the Half Width category, this value may be interpreted by saying "in 95% of repeated trials, the sample mean would be reported as within the interval sample mean ± half width."

The half width can be reduced by running the simulation for a longer period of time, and not enough replication times will lead to "insufficient" in the half width column from the simulation reports. The first half is about the mathematic equation for half width which would derive to an equation which has "n" on both sides (*). Introducing the following notation:

N = number of simulation replications

$\bar{x}$ = sample mean

s = sample standard deviation

$t_{n-1,1-\alpha/2}$ = critical value from t tables

Confidence interval: $\bar{X} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$

Half-width: h=$t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}}$ ⇒ $n = t^2{}_{n-1,1-\alpha/2} \frac{s^2}{h^2}$ (*)

The first half is about the mathematic equation for half width which would derive to an equation which has "n" on both sides (*). The second half is an approximation method to estimate the "n" which is the number of replications.

Approximation:

- Replace t by z, corresponding normal critical value
- Pretend that current "s" will hold for larger samples
- Get $n \cong z^2{}_{1-\alpha/2} \frac{s^2}{h^2}$ (where s=sample standard deviation from "initial" number n_0 of replications

Easier but different approximation:

- $n \cong n_0 \frac{h_0^2}{h^2}$ (h_0=half width from "initial" number n_0 of replications, and n grows quadratic ally as h decrease)

Experiments with Baseline Problem with M_1 =8 and M_2 =9: From initial an initial simulation run of 10 replications, 95% half-width on Ω was ±$380.89. Objective is to get that value down to ±$169.68 (2% of mean value) or less. This is done by setting:

$$n \cong 10(380.89^2/169.69^2) = 50.39 \text{ rounded up to } 51$$

Running the simulation with 100 replications (conservative based on above), Ω= 8730.49 ± 174.38, not less than 169.68, but 174.38 is still less than 2% of 8730.49(174.6089). However, in the definition of Ω, part of it is regular salary cost for ASF which is a fixed amount under same rates and hours in every experiment. Therefore, the acceptable half width has been set up to 4% of the varied amount of cost which is the total cost minus fixed cost. The initial results are shown in table 4.2, based on which the replication time is set at 1850 for the baseline problem.

Table 4.2 Initial Results for Replication Calculation

Total Cost	**Fixed Cost**	**Variable Cost**	**4% Of Variable**	**Half Width**	**Estimate Replication Time**
8602.29	7620	982.29	39.2916	119.35	1845.33

A total of 1850 replications were completed for each of the 25 experiments, for a total of 46250 observations.

4.4. Staffing Experimental Results – Baseline Problem

Fu (2002) identifies 4 main approaches for optimizing simulations:

1. Stochastic approximation (gradient-based approaches)

2. Sequential response surface methodology

3. Random search

4. Sample path optimization (also known as stochastic counterpart)

Here the approach follows the response surface methodology. The "local response surface" is used to determine a search strategy (e.g., moving to the estimated gradient direction) and the process is repeated. In other words, the meta models do not attempt to characterize the objective function in the entire solution space but rather concentrate in the local area that the search is currently exploring. The analysis of the experimental results is reported in the following sections.

4.4.1. Convexity of the Objective Function

The performance objective of the ASF operation was shown above in section 4.1. The first analysis focuses on studying the convexity behavior of this objective function. Especially the authors attempt to build an understanding of the simulation's 'response surface'. That is the combination effect of $M_{j,t}$. on the objective outcome Ω. Simulation results for three problems (#s 1, 2 and 3) are shown in Figures 4.2 to 4.7 and in Tables 4.2 to 4.4. Note the results shown the expected costs for a simulation run of 100 replications. It is clear from the expected cost curves that Ω is a strictly convex function. The 3-D response surfaces indicate though that Ω is not always smooth convex (for example problem 3).

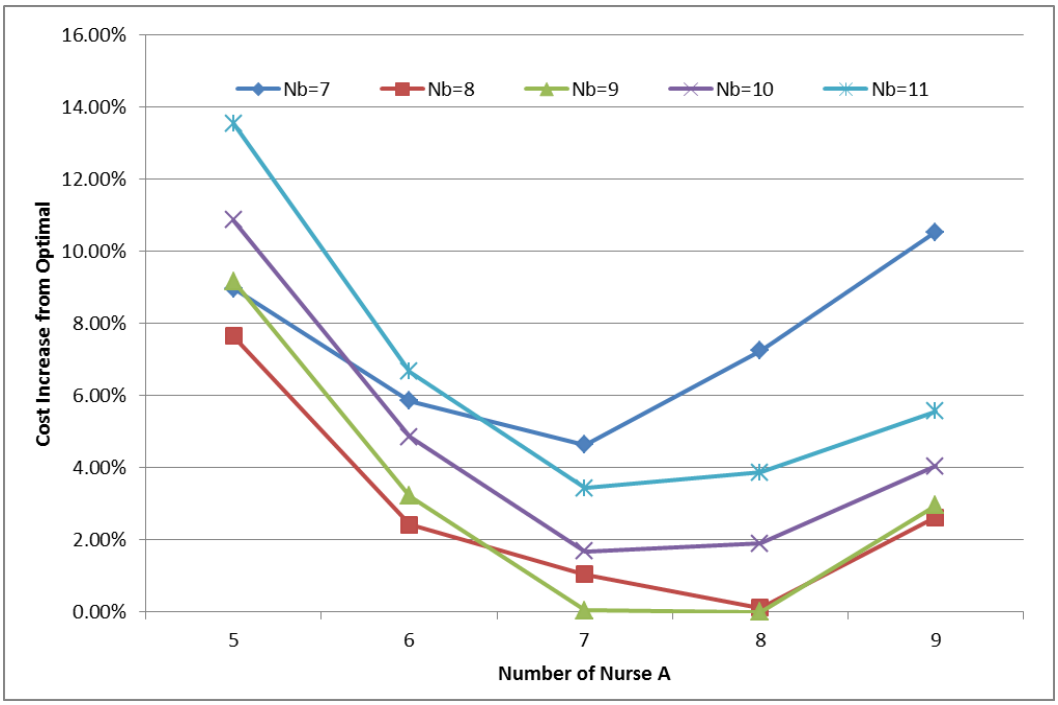


Figure 4.2 Total expected costs for problem-1.

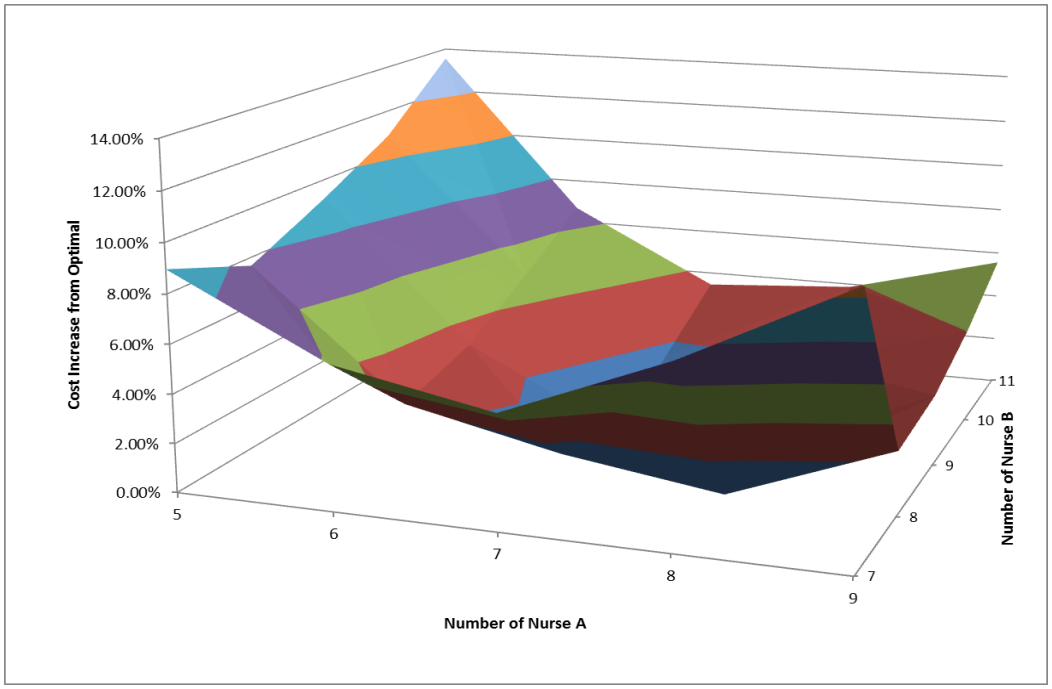


Figure 4.3 Overall performance convexities for the $N_{A\,x}\ N_B$ Decision Space – Problem 1.

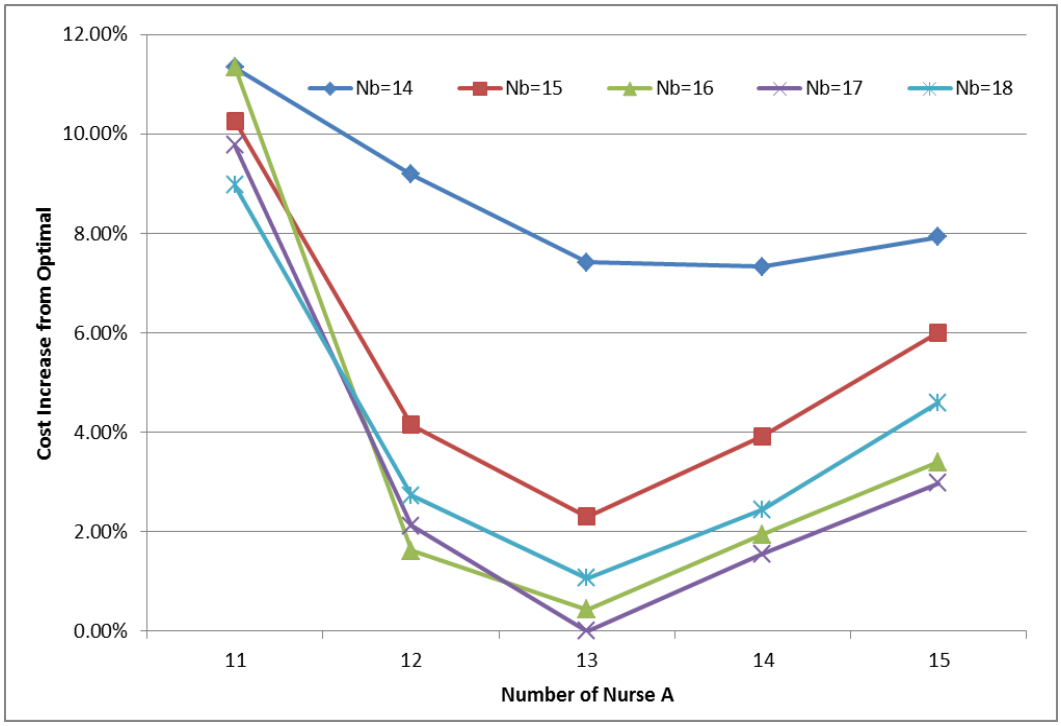


Figure 4.4 Total expected costs for problem – 2.

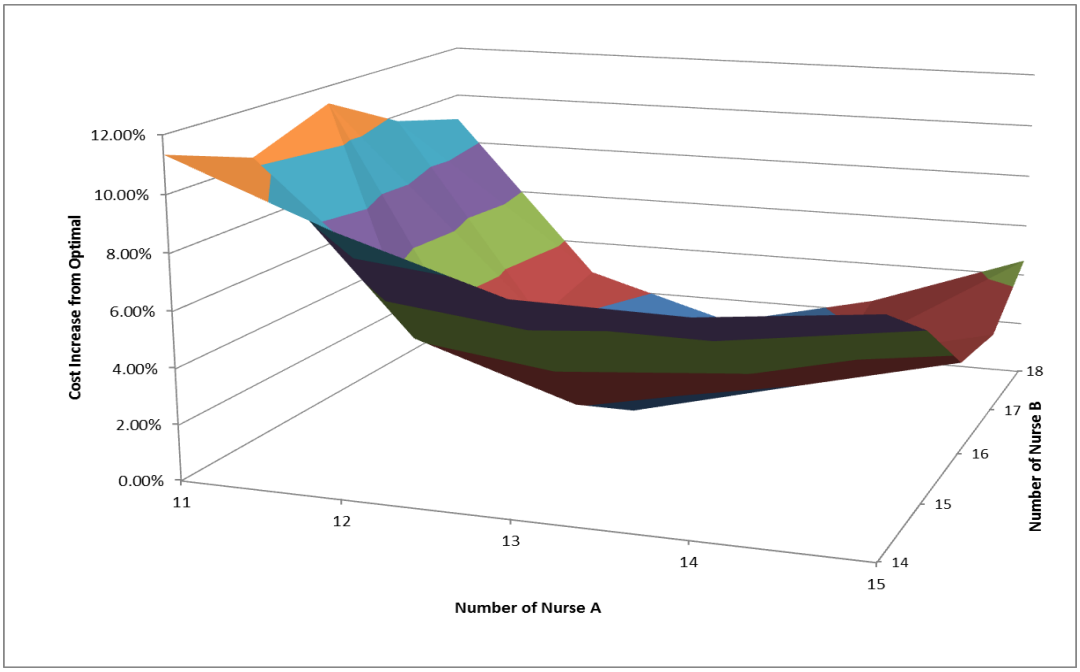


Figure 4.5 Overall performance convexity for the $N_A x N_B$ decision space – problem 2.

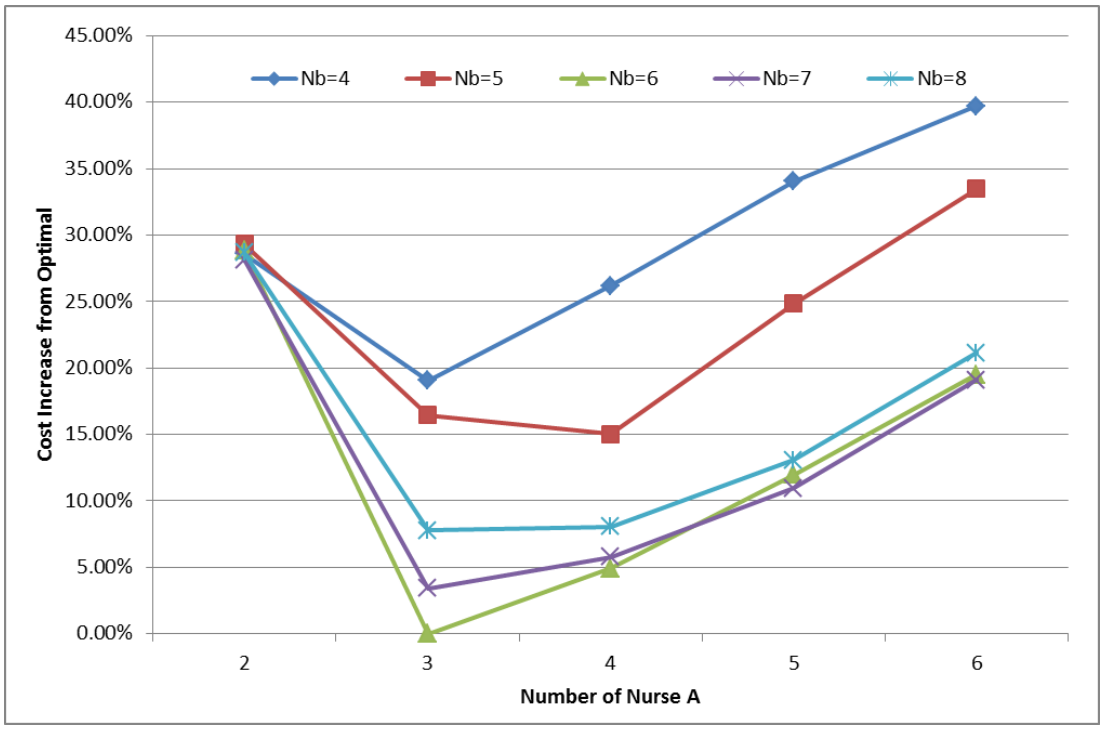


Figure 4.6 Total expected costs – problem 3.

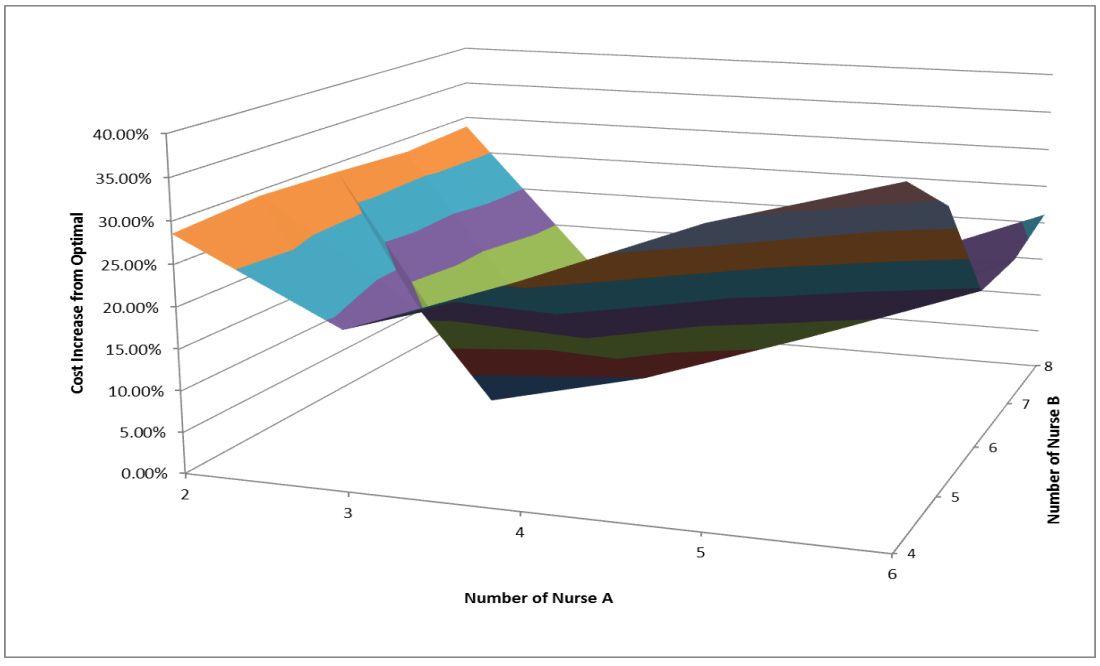


Figure 4.7 Overall performance convexities for the $N_A x N_B$ Decision Space – Problem 3.

Table 4.3 Total Expected Optimal Solution Space for Problem-1.

Total Cost ($) - Increase from Optimal				
8.97%	5.85%	4.63%	7.24%	10.53%
7.65%	2.42%	1.04%	0.12%	2.62%
9.16%	3.22%	0.05%	**0.00%**	2.95%
10.86%	4.85%	1.68%	1.90%	4.04%
13.54%	6.67%	3.44%	3.87%	5.56%

Table 4.4 Total Expected Optimal Solution Space for Problem-2.

Total Cost ($) - Increase from Optimal				
11.34%	9.19%	7.42%	7.33%	7.93%
10.26%	4.15%	2.30%	3.92%	6.00%
11.35%	1.61%	0.44%	1.94%	3.40%
9.78%	2.11%	0.00%	1.55%	2.98%
8.97%	2.72%	1.06%	2.44%	4.59%

Table 4.5 Total Expected Optimal Solution Space for Problem-3.

Total Cost ($) - Increase from Optimal				
28.57%	19.08%	26.19%	34.05%	39.71%
29.33%	16.46%	15.01%	24.83%	33.47%
28.85%	**0.00%**	4.92%	11.98%	19.50%
28.16%	3.41%	5.80%	10.96%	19.12%
28.66%	7.76%	8.08%	13.06%	21.15%

This behavior confirms that a gradient search method can efficiently be used to solve ASF problems even those with a larger number. Problem-2 is the largest problem with maximum of 28 nurses involved, and the graphs confirm that a 2-D gradient search method would work well. For ASF analysts the experiments show optimal staffing decisions can be made quickly, precluding the need for developing approximate mathematical models. For the experimental space the Ω range was for Problem-1 = 13.5%, Problem-2 = 11.3% and Problem-3 = 28.6%. Significant reductions in ASF operational costs can thus be achieved by optimizing $M_{j,t}$.

4.4.2. Robustness of Decision Space

A key issue in studying the Ω response surface is the robustness of the optimal decision (Ω*), that is the loss of optimality as the authors switch to alternate decision points. This behavior is shown in tables 4.2 to 4.4 which records (Ω-Ω*)/Ω*. In addition to Ω* the tables highlight Ω*+1% and Ω*+3% solution points. The authors observe that the robustness of the decision space is not consistent across the problems. For problem-1 the space is robust with number of solutions (Ω*+1%) = 3 and solutions (Ω*+3%) = 5. In contrast for problem-3 number of solutions (Ω*+1%) = 0 and solutions (Ω*+3%) = 0. For problem-3 the optimal solution is quite distinct and the authors observe that closest non-optimal solution is away by 3.4%. The results confirm that approximate solutions to the problem may be significantly deviant from the optimal.

4.4.3. Ω Convergence Rate

In simulation optimization a common approach is to evaluate the convergence rate of the objective function. These are derived as follows:

Nurse-A Convergence Rate $\{M_1 , M_2\} = \{\Omega(M_1+1, M_2) - \Omega(M_1, M_2)\} / \Omega^*$

Nurse-B Convergence Rate $\{M_1 , M_2\} = \{\Omega(M_1, M_2+1) - \Omega(M_1, M_2)\} / \Omega^*$

Tables 4.5 to 4.7 show the convergence rates for the three problems. The convergence rate is highest for highest Nurse-A at the lowest staffing levels. This behavior is consistent across all three problems. Problem-3 displays a non-smooth convex behavior in that the Nurse-B convergence rate does show more than one turning point. This implies a purely gradient search method may not always work in a staffing problem of this type.

Table 4.6 Ω Convergence Rate for Problem-1

	Nurse-A Convergence				
	NURSE-A				
NURSE-B	5	6	7	8	9
7		-3.35%	-1.31%	2.81%	3.53%
8		-5.62%	-1.49%	-0.99%	2.70%
9		-6.39%	-3.41%	**-0.05%**	3.17%
10		-6.45%	-3.41%	0.23%	2.31%
11		-7.39%	-3.47%	0.47%	1.82%

	Nurse-B Convergence				
	NURSE-A				
NURSE-B	5	6	7	8	9
7					
8	-1.41%	-3.69%	-3.86%	-7.66%	-8.49%
9	1.62%	0.86%	-1.06%	**-0.12%**	0.35%
10	1.82%	1.76%	1.75%	2.04%	1.17%
11	2.88%	1.95%	1.89%	2.12%	1.63%

Table 4.7 Ω Convergence Rate for Problem-2

	Nurse-A Convergence				
	NURSE-A				
NURSE-B	11	12	13	14	15
14		-2.16%	-1.77%	-0.09%	0.61%
15		-6.11%	-1.85%	1.62%	2.08%
16		-9.73%	-1.18%	1.51%	1.45%
17		-7.66%	-2.11%	1.55%	1.42%
18		-6.25%	-1.66%	1.38%	2.15%

	Nurse-B Convergence				
	NURSE-A				
NURSE-B	11	12	13	14	15
14					
15	-1.08%	-5.03%	-5.12%	-3.41%	-1.93%
16	1.09%	-2.54%	-1.86%	-1.97%	-2.60%
17	-1.57%	0.50%	-0.44%	-0.39%	-0.42%
18	-0.81%	0.61%	1.06%	0.89%	1.61%

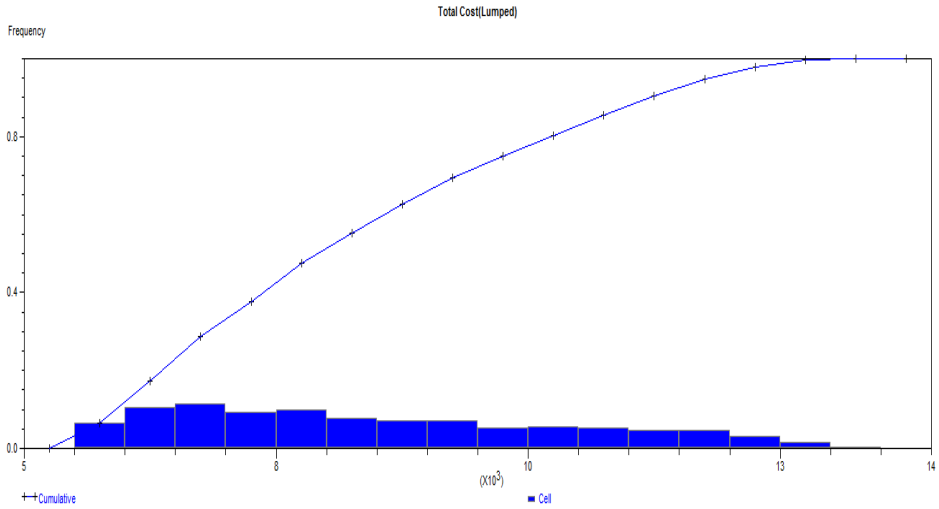
Table 4.8 Ω Convergence Rate for Problem-3

	Nurse-A Convergence				
	NURSE-A				
NURSE-B	2	3	4	5	6
4		-9.49%	7.11%	7.85%	5.66%
5		-12.87%	-1.45%	9.82%	8.63%
6		-28.85%	4.92%	7.06%	7.52%
7		-24.74%	2.38%	5.16%	8.16%
8		-20.90%	0.32%	4.98%	8.09%

	Nurse-B Convergence				
	NURSE-A				
NURSE-B	2	3	4	5	6
4					
5	0.75%	-2.62%	-11.18%	-9.21%	-6.24%
6	-0.48%	-16.46%	-10.09%	-12.85%	-13.97%
7	-0.69%	3.41%	0.88%	-1.02%	-0.38%
8	0.50%	4.34%	2.28%	2.10%	2.03%

4.5. Variance Analysis of Ω

The chart 4.2.5 displays the histogram of one experiment under 1850 replications when nurse A=8, nurse B=9, which is under enough resource level scenario but the variance range is so wide that with a high chance it may cause delays in ASFs. However, it is the natures of ASFs that the surgery variance cannot be avoid, and comparison between variance will be the next evaluate factor in next topics.



Histogram Summary
Total Cost(Lumped)

Cell	Cell Limits From	To	Abs. Freq. (Time) Cell	Cumul.	Rel. Freq. Cell	Cumul.
1	-Infinity	5500	2	2	0.001082	0.001082
2	5500	6000	120	122	0.0649	0.06598
3	6000	6500	195	317	0.1055	0.1714
4	6500	7000	213	530	0.1152	0.2866
5	7000	7500	169	699	0.0914	0.378
6	7500	8000	182	881	0.09843	0.4765
7	8000	8500	143	1024	0.07734	0.5538
8	8500	9000	132	1156	0.07139	0.6252
9	9000	9500	129	1285	0.06977	0.695
10	9500	1e+004	99	1384	0.05354	0.7485
11	1e+004	1.05e+004	101	1485	0.05462	0.8031
12	1.05e+004	1.1e+004	98	1583	0.053	0.8561
13	1.1e+004	1.15e+004	87	1670	0.04705	0.9032
14	1.15e+004	1.2e+004	84	1754	0.04543	0.9486
15	1.2e+004	1.25e+004	56	1810	0.03029	0.9789
16	1.25e+004	1.3e+004	31	1841	0.01677	0.9957
17	1.3e+004	1.35e+004	7	1848	0.003786	0.9995
18	1.35e+004	+Infinity	1	1849	0.0005408	1

Figure 4.2.5 Histogram for A8B9 with 1850 replications.

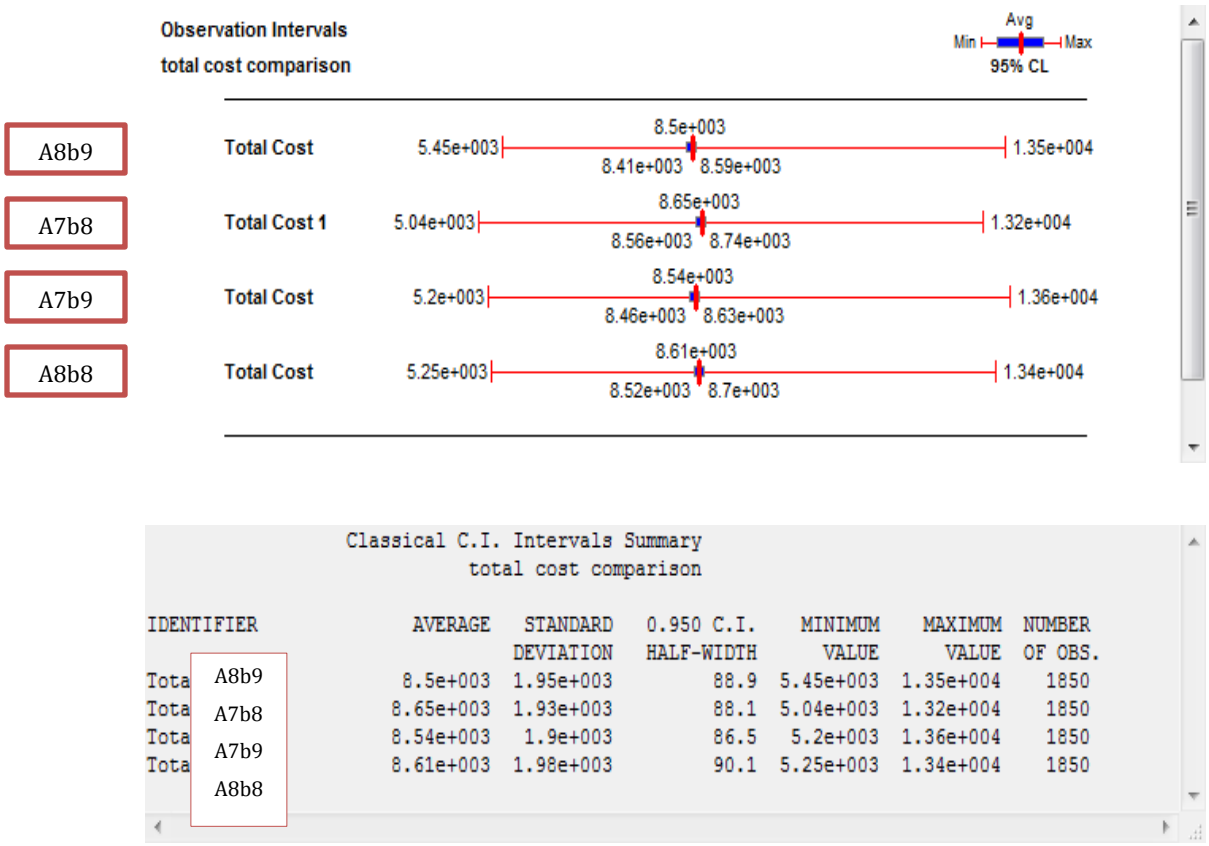


Figure 4.2.6 A8b9, a8b8, a7b9, and a7b8 total cost comparison.

	Doc delay	Staff overflow	Patient delay	Regular salary	Total cost
A7b8(1)	2687.69	597.37	537.84	4829.00	8651.90
A7b9(2)	2458.59	551.97	506.87	5027.00	8544.44
A8b9(3)	2220.50	497.23	482.78	5302.00	8502.52
A8b8(4)	2442.66	546.98	514.47	5104.00	8608.11

Figure 4.2.7 Separate mean cost for A8b9, a8b8, a7b9,

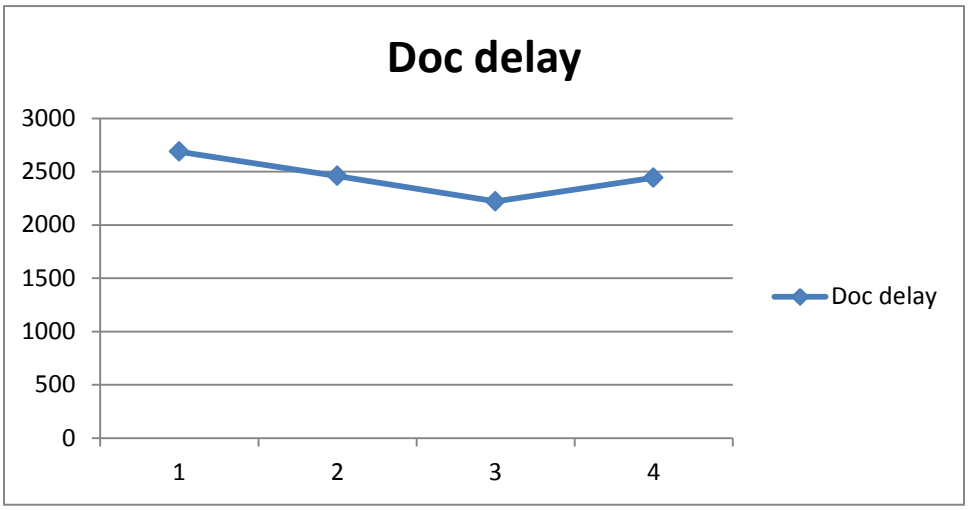


Figure 4.3.1 Doctor's delay and staff's overflow mean cost for A8b9, a8b8, a7b9, a8b8.

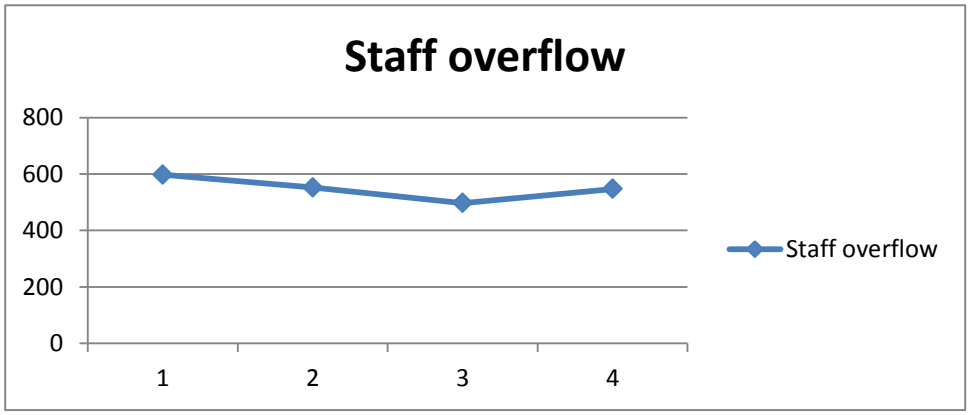


Figure 4.3.2 Staff delay and staff's overflow mean cost for A8b9, a8b8, a7b9, a8b8.

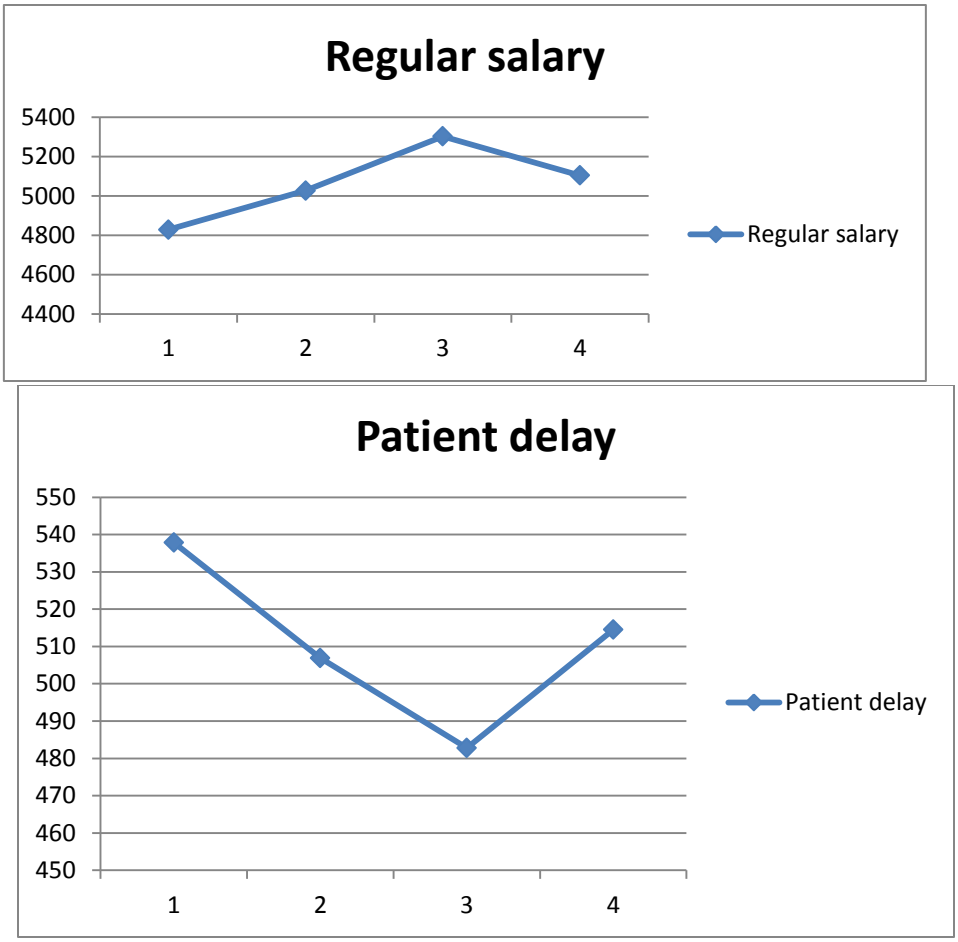


Figure 4.3.3 Patients' delay and staff's overflow mean cost for A8b9, a8b8, a7b9, a7b8.

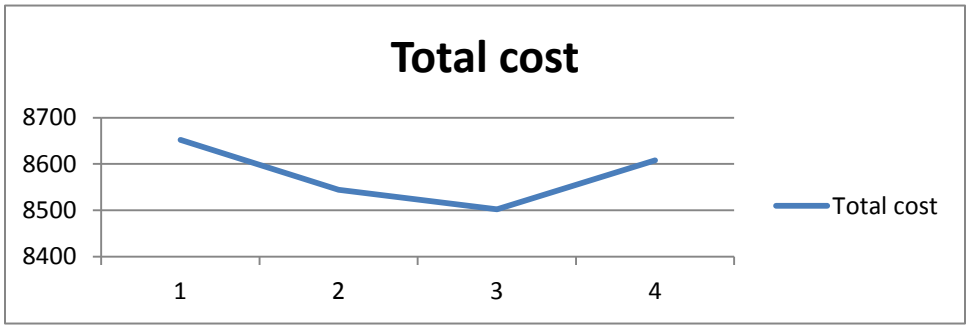


Figure 4.3.4 Total mean cost for A8b9, a8b8, a7b9, a7b8.

Though in a small range within that flat area which is at the bottom of the U-convex graph, same growing and decreasing trend is in those four factors has also been displayed here through these five graphs: as the number of staffing increase, the doctor's delay penalty goes down so as staffing's overflow penalty and the patient delay penalty, however, at the point of nurse A=8. Nurse B=9, it reaches the lowest total cost under the highest staffing salary, as you can predict that even more staffing have been added into the system, higher regular salary would prevent it to be the lowest total cost, in the other words, when nurse A=8, Nurse B=9 is the optimal result even in that equal good area under specific conditions.

4.6. Sensitivity Analysis of Physician Delay Penalty

Next the authors study the sensitivity effect of the physician delay penalty (ϕ_D) on the decision space. Noting that this is a key feature of the proposed objective function, ASF operators which to learn more about how the decision space is effected for decreasing and increasing values of ϕ_D . Experimental results are shown in Tables 4.12 to 4.14.

Table 4.12 Decision Sensitivity to Changing ϕD –Problem 1

ϕ_D	M_1	M_2	Ω	Staff Cost
$150	6	8	$6,048	$3,886
$200	7	8	$6,556	$4,107
$250	7	9	$6,987	$4,267
$300	8	9	$7,421	$4,511
$350	8	9	$7,820	$4,511
$400	8	9	$8,219	$4,511
$450	8	9	$8,618	$4,511

Table 4.13 Decision Sensitivity to Changing ϕD –Problem 2

ϕ_D	M_1	M_2	Ω	Staff Cost
$150	12	15	$9,941	$8,143
$200	12	16	$10,277	$8,363
$250	13	16	$10,567	$8,608
$300	13	17	$10,834	$8,812
$350	13	17	$11,110	$8,812
$400	13	17	$11,388	$8,812
$450	13	17	$11,665	$8,812

Table 4.14 Decision Sensitivity to Changing ϕD –Problem 3

ϕ_D	M_1	M_2	Ω	Staff Cost
$150	3	5	$3,566	$2,475
$200	3	6	$3,628	$2,605
$250	3	6	$3,789	$2,605
$300	3	6	$3,960	$2,605
$350	3	6	$4,131	$2,605
$400	3	6	$4,303	$2,605
$450	3	6	$4,474	$2,605

CHAPTER 5

ASSIGNMENT OF SCHEDULE BLOCKS TO PHYSICIAN GROUPS

In chapter 3 the authors introduced the block scheduling arrangement that ASFs negotiate with physician groups. Since many schedule combinations are possible, In this chapter the authors develop and evaluate heuristic assignment rules to derive block schedules which optimize performance. Currently such assignments are done manually by ASF managers using their past experiences. The heuristics developed here would provide analytical solutions with the capability to handle relatively large problems.

5.1. Defining the Physician Block Assignment Problem

An ASF may have flexibility in the way it assigns schedule blocks to the different physician groups that are active in the ASF. A schedule block is defined as a continuous window, usually 3 to 4 hours long, during which assigned physician groups can schedule their surgery patients. A physician group k will contract with the ASF for to perform surgeries during E_k continuous blocks. Typically, E_k is derived from the capacity requirements of the group and here the authors limit E_k to an integer. The assumption here is that the assignment of blocks to physicians effects the overall performance Ω of the ASF. There a multiple reasons for this including the surgery types, surgery time variances, resources requirements. When $E_k > 1$ and physician group k performs multiple

surgery types, the patients must also be divided among the assigned blocks. The decision variables then are:

$\delta_{k,t}$ Physician group k is assigned schedule block t (1 = yes, 0 = no)

$A_{i,t}$ Number of patient type i scheduled to arrive in block t

The ASF physician block assignment problem is then described as determining $\delta_{k,t}$ such that the expected value of Ω is minimized. Where the decision space is constrained such that $\sum \delta_{k,tt} = E_k$ and $A_{i,t} = 0$ if $\delta_{k,t} = 0$ and $\delta_{k,t} = 1$. Further the authors assume the staffing level has already been fixed, that is $M_{j,t}$ is predetermined. The ASF operational objective has been previously defined as follows:

$$Minimize: \quad \Omega = \sum_j \sum_t \left(4\varphi_{j,R} M_{j,t}\right) + \sum_j \left(\varphi_{j,O} O_j\right) + \phi_P T_P + \phi_D T_D$$

Since $M_{j,t}$ is fixed then the first term in Ω is a constant for a given problem. The effect of $\delta_{k,t}$ on the objective function variables T_P and T_D is determined from the ASF simulation model, and used to evaluate the quality of the decision policy. Table 5.1. shows an example decision policy.

Table 5.1. An Example Physician Assignment Decision Policy

Physician Group k	Blocks >>	t=1	t=2	t=3
1	$\delta_{1,t}$	1	1	0
	$A_{1,t}$	4	1	0
	$A_{2,t}$	0	4	0
	$A_{3,t}$	2	2	0
2	$\delta_{2,t}$	1	0	1
	$A_{4,t}$	5	0	2
	$A_{5,t}$	0	0	5
3	$\delta_{3,t}$	0	0	1
	$A_{6,t}$	0	0	5

The research strategy is to leverage classical machine sequencing algorithmic knowledge to develop several heuristics for the determination of $\delta_{k,t}$. These heuristics are then tested using the ASF simulation model to characterize their performance. All heuristics first determine $\delta_{k,t}$ and then $A_{i,t}$. In a basic heuristic $A_{i,t}$ is determined using the load balanced surgery schedule (section 3.7) while in an extended heuristic additional rules for the derivation of $A_{i,t}$ are introduced.

5.2. Similarity from Machine Scheduling

The three main topics in machine scheduling are single or parallel machine sequencing, flow shop sequencing and job shop scheduling. Since the definition of scheduling almost

covered sequencing, though they focused on different aspects, the scheduling is chosen to stand for sequencing and scheduling in the following content. Single-machine scheduling or single-resource scheduling is the process of assigning a group of tasks to a single machine or resource. The tasks are arranged so that one or many performance measures may be optimized. Parallel machines are parallel identical machines meaning that tasks or jobs can be finished by either of the machines. The main difference between single machine sequencing and flow shop sequencing is that more machine quantities and given process order (the definition of flow shop scheduling is given later). However, the range of job shop scheduling is wider than that of flow shop scheduling, for example, both with process orders, usually one job is not allowed to rework in the same machine in the flow shop scheduling problems but there is no path route rule for jobs in job shop scheduling problems.

With about 70 years' investigation, major findings include: Graham had already provided the List scheduling algorithm in 1966, which is $(2 - 1/m)$-competitive, where m is the number of machines.[1] Also, it was proved that List scheduling is optimum online algorithm for 2 and 3 machines. The Coffman–Graham algorithm (1972) for uniform-length jobs is also optimum for two machines, and is $(2 - 2/m)$-competitive.[2][3] In 1992, Bartal, Fiat, Karloff and Vohra presented an algorithm that is 1.986 competitive.[4] A 1.945-competitive algorithm was presented by Karger, Philips and Torng in 1994.[5] In 1992, Albers provided a different algorithm that is 1.923-competitive.[6] Currently, the best known result is an algorithm given by Fleischer and Wahl, which achieves a competitive ratio of 1.9201.[7]A lower bound of 1.852 was presented by Albers.[8] Taillard instances has an important role in developing job shop scheduling with

makespan objective.In 1976 Garey provided a proof[9] that this problem is NP-complete for m>2, that is, no optimal solution can be computed in polynomial time for three or more machines (unless P=NP).

By looking back to our ASF physician scheduling problems, patients are like jobs going through three processes (pre-operating, operating and post-operating process) with fixed order and will be helped by resources staffing members and physician groups using specific facilities. Therefore, the ASF physician scheduling problem is more like a flow shop scheduling problem.

There is a long history in time that people have devoted on best algorithms for different flow shop situations. From the Wikipedia, the Flow Shop Scheduling Problems, or FSPs, are a class of scheduling problems with a work shop or group shop in which the flow control shall enable an appropriate sequencing for each job and for processing on a set of machines or with other resources 1,2,...,m in compliance with given processing orders. Especially the maintaining of a continuous flow of processing tasks is desired with a minimum of idle time and a minimum of waiting time. FSP may apply as well to production facilities as to computing designs. In a short word, the FSP is about to schedule some jobs or tasks on machines or resources to reach some specific performance objectives like the makespan, total completion time and so on. To minimize makespan, a heuristic algorithm by S.M. Johnson can be used to solve the case of a 2 machine N job problem when all jobs are to be processed in the same order but with 3 or more machines, it may not be the optimal.

In the case like ASF, physician groups scheduling problems are like class of scheduling problems with a management work shop in which the flow control shall enable an appropriate sequencing for each patient and for processing pre- operating, operating and post-operating on set of resources including staffing members, facilities and physician groups in compliance with given processing order (as it known to all from pre-operating to operating to post-operating), and the performance objective is to reduce the idle time for doctors, reduce the overflow time for the staffing members and reduce the waiting time for patients. Accordingly, these objectives in ASF physician scheduling problems are matching different terms in FSP (will be explained in details in later sections) and the set-up of objective function considering multi aspect requirement has already been mentioned in the previous chapter three. Within certain complexities, there is no optimal solution for physician group scheduling problem and the advantages of discrete-event simulation has been explained in details in previous chapter, the following sub-section is about original problem extension and simulation model related assumptions.

5.3. General Assumptions

- The ASF opens at least half-hour before the first block for patient registrations
- The ASF remain open after the last block in overtime mode till all patients are processed
- No splitting of processes is allowed

- Patient arrivals are independent

Several heuristics were developed and tested. The authors present here five of the most promising heuristics. All of these uses concepts and approaches utilized in the machine scheduling and sequencing literature. Steps 1 to 4 are common for all the heuristics.

5.4. Heuristic #1 – Resource Balancing Algorithm

The objective of this heuristic is to generate an assignment which minimizes the imbalance in staff resource usage between the blocks. Balancing algorithms are widely used in the scheduling literature and the authors follow the same approach here.

Step – 1: Calculate the total resource usage by patient type i. Let $T_{i,j}$ be the robust estimate of the total use of resource j by type i, derived as follows:

$$T_{i,j} = \alpha_i \left\{ \sum_{n=1}^{2} \chi_{i,n,j}\left(\mu_{i,n} + z_{0.75}\sigma_{i,n}\right) \;+\; \chi_{i,3,j}\left(0.2\mu_{i,n} + z_{0.75}\sigma_{i,n}\right) \right\}$$

To derive a robust estimate the authors use the classical theory of constraints approach by adding a time buffer to each critical task. Since each resource is critical in this case, the authors add the buffer such that the actual activity time is 75% likely to be less than our robust estimate.

Step – 2: Calculate the perfect balance resource usage level for each staffing resource as follows:

$$\upsilon_j = \frac{1}{B}\sum_i T_{i,j}$$

Step – 3: Calculate M^*_j the minimum staffing level needed to meet the patient requirements plus a downtime buffer. The down time buffer ζ_j for a staffing resource accounts for the inherent continuity gap between surgeries (5+%) plus the normal staff rest time (10+%).

$$M_j^* = INT\left\{\left(1+\varsigma_j\right)\frac{\upsilon_j}{240}\right\}$$

INT is a function which returns the next largest integer. Each schedule block is assumed to be 240 minutes long without loss of generality. Here the authors assume ζ_j is a management decision based on location specific work policies and combination of surgeries. But ζ_j could be a variable that is also investigated through the simulation experimentation process. For example in a location where a larger number of short surgeries are performed then the continuity gap buffer tends to be smaller and ζ_j also smaller.

Note that any solution with a non-zero M_j is a feasible solution to the ASF physician assignment problem, for example M_j =1 for all j is feasible but will result in an excessively large Ω. Then M^*_j represents the baseline feasible solution and is used to compare the performance of the different heuristics by keeping the staffing cost constant at this level.

Step – 4: Calculate the staff resource usage requirements for each physician group as follows:

$$\Upsilon_{k,j} = \sum_{i} T_{i,j} \mid \beta_{i,k} = 1$$

Step – 5: Formulate the physician assignment balanced staffing resource load problem as a mixed integer program. The objective function in this program minimizes the total absolute resource usage variance from the perfect balance level

Objective:

$$Minimize \quad \sum_{t}\sum_{j} \omega_{j,t}$$

Such that:

$$V_{j,t} = \sum_{k} \delta_{k,t} \frac{\Upsilon_{k,j}}{E_k} \qquad for\ all\ j\ and\ t$$

$$\omega_{j,t} \geq V_{j,t} - \upsilon_j \qquad for\ all\ j\ and\ t$$

$$\omega_{j,t} \geq \upsilon_j - V_{j,t} \qquad for\ all\ j\ and\ t$$

$$\sum_{t} \delta_{k,t} = E_k \qquad for\ all\ t$$

Where:

$\delta_{k,t}$ = (0,1) and $\omega_{k,t} \geq 0$ for all k and t

$\delta_{k,t}$ = 1 for all t when E_k = 3

$\delta_{k,2}$ = 1 when E_k = 2

Step – 6: Formulate the patient assignment to physician block for balanced staffing usage problem as a linear program. The objective function in this program minimizes the total absolute resource usage variance from the perfect balance level.

Objective:

$$Minimize \quad \sum_{t} \sum_{j} \Theta_{j,t}$$

Such that:

$$\Theta_{j,t} \geq \sum_{i} A_{i,t}\, T_{i,j} \; - \; \upsilon_j \qquad for\ all\ j\ and\ t$$

$$\Theta_{j,t} \geq \upsilon_j \; - \; \sum_{i} A_{i,t}\, T_{i,j} \qquad for\ all\ j\ and\ t$$

$$\sum_{t} A_{i,t} = \alpha_i \qquad for\ all\ i$$

$$A_{i,t} \leq \; \delta_{k,t}\, \alpha_i \qquad for\ all\ i\ and\ t\ where\ \beta_{i,k} = 1$$

$$\sum_{i} \beta_{i,k}\, A_{i,t} \geq 0.9 \sum_{i} \frac{\beta_{i,k}\, \alpha_i}{E_k} \qquad for\ all\ k\ and\ t$$

Where:

$A_{i,t} \geq 0$ for all i and t

$\Theta_{j,t} \geq 0$ for all j and t

The last constraint limits the imbalance between blocks for a specific physician group to 10%. In the absence of these constraints, the solution could generate high levels of imbalance which are infeasible for the group.

5.4. Heuristic #2 – Asymmetrical Resource Balancing Algorithm

In heuristic #1 the objective was to balance the resources load across all the scheduling windows. A detailed review of the generated solutions provides specific insights into possible improvement strategies. A key observation was that due to the uneven resource requirements between physician groups, there is an inherent imbalance in the solution. Solutions where the higher loaded assignments are in the first or second window, that is $V_{j,1} > v_j$, tend to outperform the inverse solutions where $V_{j,1} < v_j$. The front loaded solutions tend to opportunistically utilize the slack in the system, and thus have lower levels of overtime. Interestingly, this rule was mentioned in the discussions with ASF administrators. The objective of heuristic #2 is to generate an asymmetrical resource balance, that is the average loading of all resources is higher in block $t=1$ compared to $t=3$.

Step – 1-4: Same as Heuristic #1

Step – 5: Formulate the physician assignment balanced staffing resource load problem as a mixed integer program. The objective function in this program minimizes the total absolute resource usage variance from the target asymmetrical balance level . Introducing:

W_t Asymmetry rate (-25% to +25%) by which block *t* target load is offset from perfect balance

Here the authors set W_1=0.1, W_2=0, and W_3=-0.1. Observe that W_t=0 indicates no asymmetry for that period. In heuristic #1 the objective gives equal importance to all the staffing resources. In this heuristic the objective is expanded such that the priority of each resource is weighted by the ratio of their regular plus overtime cost rates to the average rates for all resources. The MIP is then formulated as follows:

Objective:

$$Minimize \quad \frac{1}{\frac{1}{3}\sum_j \varphi_{j,R} + \varphi_{j,O}} \sum_t \sum_j \left(\varphi_{j,R} + \varphi_{j,O}\right) \omega_{j,t}$$

Such that:

$$V_{j,t} = \sum_k \delta_{k,t} \frac{\Upsilon_{k,j}}{E_k} \qquad for\ all\ j\ and\ t$$

$$\omega_{j,t} \geq V_{j,t} - (1 + W_t)v_j \qquad for\ all\ j\ and\ t$$

$$\omega_{j,t} \geq (1 + W_t)v_j - V_{j,t} \qquad for\ all\ j\ and\ t$$

$$\sum_t \delta_{k,t} = E_k \qquad for\ all\ t$$

Where:

$\delta_{k,t}$ = (0,1) and $\omega_{k,t} \geq 0$ for all *k* and *t*

$\delta_{k,t}$ = 1 for all *t* when E_k = 3

$\delta_{k,2}$ = 1 when E_k = 2

Step – 6: Also adding the weighted parameter for the objective formula.

Objective:

$$Minimize \quad \frac{1}{\frac{1}{3}\sum_j \varphi_{j,R} + \varphi_{j,O}} \sum_t \sum_j \left(\varphi_{j,R} + \varphi_{j,O}\right) \Theta_{j,t}$$

Such that:

$$\Theta_{j,t} \geq \sum_i A_{i,t}\, T_{i,j} \; - \; (1 + W_t) v_j \qquad for\ all\ j\ and\ t$$

$$\Theta_{j,t} \geq \; (1 + W_t) v_j \; - \; \sum_i A_{i,t}\, T_{i,j} \qquad for\ all\ j\ and\ t$$

$$\sum_t A_{i,t} = \alpha_i \qquad for\ all\ i$$

$$A_{i,t} \leq \; \delta_{k,t}\, \alpha_i \qquad for\ all\ i\ and\ t\ where\ \beta_{i,k} = 1$$

$$\sum_i \beta_{i,k}\, A_{i,t} \geq 0.9 \sum_i \frac{\beta_{i,k}\, \alpha_i}{E_k} \qquad for\ all\ k\ and\ t$$

Where:

$A_{i,t} \geq 0$ for all i and t

$\Theta_{j,t} \geq 0$ for all j and t

5.4. Heuristic #3 – Pre-operative and Resource Balancing Algorithm

From a review of the solutions for the previous problems, the authors find that in some cases the generated solution while having staffing resource use is imbalanced at the activity level. That is total activity times for pre-operation, surgery or PACU are not

balanced. The result is that the simulation results show weak performance for these schedules, because longer activity queues tend to form during blocks with high activity use. In this heuristic the authors attempt to address this issue by adding an additional constraint to Step #6, which limits the imbalance in pre-operation activity time to ±10%. The authors also found that attempting to balance all activities simultaneously was not an effective approach, since the solution space is overtly restricted.

Step – 1-5: Same as Heuristic #2

Step – 6: Formulate the patient assignment to physician block problem for balancing both staffing usage and pre-operative activity times. The objective function in this program minimizes the total absolute resource usage variance from the weighted balance level.

Objective:

$$Minimize \quad \frac{1}{\frac{1}{3}\sum_j \varphi_{j,R} + \varphi_{j,O}} \sum_t \sum_j \left(\varphi_{j,R} + \varphi_{j,O}\right) \Theta_{j,t}$$

Such that:

$$\Theta_{j,t} \geq \sum_i A_{i,t}\, T_{i,j} \; - \; (1 + W_t) v_j \qquad for\ all\ j\ and\ t$$

$$\Theta_{j,t} \geq \; (1 + W_t) v_j \; - \; \sum_i A_{i,t}\, T_{i,j} \qquad for\ all\ j\ and\ t$$

$$\sum_t A_{i,t} = \alpha_i \qquad for\ all\ i$$

$$A_{i,t} \leq \; \delta_{k,t}\, \alpha_i \qquad for\ all\ i\ and\ t\ where\ \beta_{i,k} = 1$$

$$\sum_i \beta_{i,k} A_{i,t} \geq 0.85 \sum_i \frac{\beta_{i,k} \alpha_i}{E_k} \qquad for\ all\ k\ and\ t$$

$$\sum_i A_{i,t}(\mu_{i,1} + z_{0.75}\sigma_{i,1}) \geq \frac{0.90}{3} \sum_i \alpha_i(\mu_{i,1} + z_{0.75}\sigma_{i,1}) \quad for\ all\ i\ and\ t$$

Where:

$A_{i,t} \geq 0$ for all i and t

$\Theta_{j,t} \geq 0$ for all j and t

In balancing the pre-operation time the minimum level is set at 90% for the 3 schedule blocks. A robust activity processing time estimate (75% likelihood) is used.

5.4. Heuristic #4 – Operative and Resource Balancing Algorithm

This heuristic is similar to Heuristic #3 in that it attempts to also balance the total processing time for a specific activity. Here the focus activity is the surgery processing time. This is a key activity in that not only does it affect resource usage but it has a direct impact on physician delay a significant component of the objective function.

Step – 1-5: Same as Heuristic #2

Step – 6: Formulate the patient assignment to physician block for balanced both staffing usage and operative time problem as a linear program. The objective function in this program minimizes the total absolute resource usage variance from the weighted balance level.

Objective:

$$Minimize \quad \frac{1}{\frac{1}{3}\sum_j \varphi_{j,R} + \varphi_{j,O}} \sum_t \sum_j \left(\varphi_{j,R} + \varphi_{j,O}\right) \Theta_{j,t}$$

Such that:

$$\Theta_{j,t} \geq \sum_i A_{i,t}\, T_{i,j} \; - \; (1 + W_t) v_j \qquad for\ all\ j\ and\ t$$

$$\Theta_{j,t} \geq \; (1 + W_t) v_j \; - \; \sum_i A_{i,t}\, T_{i,j} \qquad for\ all\ j\ and\ t$$

$$\sum_t A_{i,t} = \alpha_i \qquad for\ all\ i$$

$$A_{i,t} \leq \; \delta_{k,t}\, \alpha_i \qquad for\ all\ i\ and\ t\ where\ \beta_{i,k} = 1$$

$$\sum_i \beta_{i,k}\, A_{i,t} \geq 0.85 \sum_i \frac{\beta_{i,k}\, \alpha_i}{E_k} \qquad for\ all\ k\ and\ t$$

$$\sum_i A_{i,t}\left(\mu_{i,2} + z_{0.75}\sigma_{i,2}\right) \geq \frac{0.90}{3} \sum_i \alpha_i \left(\mu_{i,2} + z_{0.75}\sigma_{i,2}\right) \quad for\ all\ i\ and\ t$$

Where:

$$A_{i,t} \geq 0 \text{ for all } i \text{ and } t$$

$$\Theta_{j,t} \geq 0 \text{ for all } j \text{ and } t$$

In balancing the surgery activity processing time the minimum level is set at 90% for the 3 schedule blocks. A robust activity processing time estimate (75% likelihood) is used.

5.4. Heuristic #5 – Priority Blocks and Balancing Algorithm

In balancing the surgery activity processing time the minimum level is set at 90% for the 3 schedule blocks. A robust activity processing time estimate (75% likelihood) is used.

Step – 1-4: Same as Heuristic #1

Step – 5: Block arrangement is based on *ScoreK*, the highest *ScoreK* will be arranged first in the early time slot. When the first V_j is greater than v_j, go to next higher *ScoreK*'s physician group and assign the next available block.

$$Score\ K = \frac{\sum_j W_{t,j} \times \Upsilon_{k,j}}{E_k} \quad for\ all\ t\ and\ set\ the\ higher\ group\ k\ \delta_{k,t} = 1\ until$$

$$V_{j,t} > v_j \quad for\ all\ t$$

$$where\ V_{j,t} = \sum_k \delta_{k,t} \frac{\Upsilon_{k,j}}{E_k} \qquad for\ all\ j\ and\ t$$

$$v_j = \frac{1}{B} \sum_i T_{i,j}$$

Step – 6: Same as Heuristic #2

5.5. Test Problems for Heuristics Evaluation

The evaluation plan was to generate the physician assignment solution ($\square_{k,t}$ and $A_{i,t}$) for a diverse set of problems, and then generate the performance measure Ω using the ASF simulator developed in chapter 3. Classical approaches to heuristic research require a comparative analysis of candidate heuristics across a range of problems. Using the set of surgeries introduced in chapter 3 and the associated parameters, a set of 10 benchmark test problems were designed. Key attributes of the problems are shown in table 5.3. We have used concepts and approaches used in classical assembly line balancing to develop these problems.

Table 5.3 Set of Benchmark Test Problems

Problem #	Patient Types i	Total Arrivals $\sum_i \alpha_i$	Physician Groups H	Total Physicians $\sum_k N_k$	Blocks / Group $Avg\ E_k$	Nurse-A Loading (j=1)	Nurse-B Loading (j=2)	MedTech Loading (j=3)
1	20	144	5	11	2.00	75%	72%	72%
2	30	274	6	18	2.33	85%	80%	80%
3	15	116	4	9	2.25	80%	75%	80%
4	30	350	6	22	2.83	82%	77%	77%
5	30	286	6	22	2.33	82%	78%	77%
6	16	78	4	9	1.50	85%	75%	75%
7	19	140	4	9	2.75	80%	70%	74%
8	21	133	5	11	1.80	85%	73%	70%
9	18	36	6	6	1.00	70%	60%	70%
10	21	74	5	9	1.40	70%	75%	79%

Problems #2, #4 and #5 are relatively large problems with 30 patient types, with problem #4 being the largest with 350 patients/day. In contrast problem #6 is the

smallest problem with only 16 patient types and serving only 78 patients/day. Problem #10 is a highly diverse problem in that its 74 patient/day are distributed over 21 patient types implying a large variety of resource use profiles are in play. Problems #8, #9 and #10 have an average E_k<2 implying a large decision space for the physician assignment problem. In contrast, problems #4 and #7 have an average E_k>2.75 implying a small; decision space.

Problems # 1 and #3 are nominal problems in hat most of their descriptive metrics are at a mean level. From table 5.3 the authors that resource loading levels range from 60% to 85% with most problems having a loading in the 70 to 80% range. The loading level was kept in a narrow range by design to minimize the effect of surplus staffing capacity on the heuristics evaluation process.

5.5.1. Replication Estimate for the Experiments

Similar to the analysis done in section 4.3.2, a series of initial simulation experiments were conducted to derive the valid replication number for each of the test problems. As noted earlier simulation experiments are inherently characterized by errors or measure variance. For a valid study the simulation replication number should be estimated to get more accurate experimental results. Table 5.4 shows the results for the test problems using an initial run of 200 replications.

For each problem the staffing level is constant across all the heuristics being evaluated, thus the staffing cost is fixed. The variable cost then included the patient delay and physician delay penalties plus the overtime costs. The variable cost is of primary interest here, and is thus separated out from the total cost. Based on the half width data

the replication number for each problem was calculated and reported in table 5.5. Arena's built in "Output Analyzer" offers us a convenient function of doing outputs analysis and this tool was utilized here.

Table 5.4 Statistical behavior of test problems (200 Replications)

Problem	Total Cost	Fixed cost	Variable Cost	4% of Variable	Half Width
1	8602.29	$7,620	$982	39.29	119.35
2	$15,927	$13,764	$2,162	86.51	206.84
3	$6,553	$5,568	$984	39.38	96.88
4	$19,727	$18,108	$1,619	64.78	131.38
5	$16,256	$14,892	$1,364	54.57	145.58
6	$5,049	$3,876	$1,173	46.93	143.21
7	$7,853	$6,960	$893	35.74	118.99
8	$7,742	$6,564	$1,178	47.14	149.81
9	$2,903	$2,088	$815	32.61	93.63
10	$4,981	$3,828	$1,153	46.10	126.61

Table 5.5 Simulation Replication for Test Problems

Problem	1	2	3	4	5	6	7	8	9	10
Reps	1850	2120	1790	1140	1430	2440	3670	2020	1650	1920

5.6. Decisions Generated by the Heuristics

For all 10 problems the decision or solution sets ($\delta_{k,t}$ and $A_{i,t}$) were first generated by the 5 heuristics. For problem #1 the results are shown in Table 5.6. Reviewing first $\delta_{k,t}$ the authors note that by design heuristics #2, #3 and #4 have the same decisions. Comparing heuristics #1 and #2 the authors that apart from physician group k=4, the two solutions are completely different. Clearly, the move to an asymmetrical balance had an impact on the decision policy. Reviewing the $A_{i,t}$ decisions the authors compare the results of heuristics #2, #3 and #4 since they have the same $\delta_{k,t}$ decision. Apart from group k=3 and k=5, the authors see that the decisions vary significantly for the other groups across these three heuristics. Heuristic #3 tends to concentrate same surgery types into the same block. For example in the case of k=2 the authors see that patients with the same surgery type are scheduling in the same block. Heuristic #4 on the other hand is closer to the #2 solution, clearly surgery processing time balance is achieved by smaller changes relative to the #2 solution.

Table 5.6 Heuristic Decisions – Problem #1

Group k	*Blocks* t	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
		Heuristic #1			**Heuristic #2**			**Heuristic #3**			**Heuristic #4**			**Heuristic #5**		
1	$\delta_{1,t}$	0	1	1	1	1	0	1	1	0	1	1	0	1	1	0
	$A_{9,t}$	0	5	0	0	5	0	0	5	0	0	5	0	3	2	0
	$A_{10,t}$	0	3	2	4	1	0	0	5	0	2	3	0	0	5	0
	$A_{19,t}$	0	0	4	4	0	0	4	0	0	2	2	0	4	0	0
	$A_{20,t}$	0	0	4	2	2	0	4	0	0	4	0	0	3	1	0
2	$\delta_{2,t}$	1	1	0	0	1	1	0	1	1	0	1	1	1	1	0
	$A_{7,t}$	3	3	0	0	6	0	0	6	0	0	5	1	6	0	0
	$A_{8,t}$	6	0	0	0	3	3	0	0	6	0	0	6	1	5	0
	$A_{17,t}$	0	5	0	0	0	5	0	0	5	0	1	4	1	4	0
	$A_{18,t}$	2	1	0	0	0	3	0	3	0	0	3	0	3	0	0
3	$\delta_{3,t}$	0	0	1	1	0	0	1	0	0	1	0	0	1	0	0
	$A_{6,t}$	0	0	3	3	0	0	3	0	0	3	0	0	3	0	0
	$A_{6,t}$	0	0	3	3	0	0	3	0	0	3	0	0	3	0	0
	$A_{6,t}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	$A_{6,t}$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	$\delta_{4,t}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	$A_{6,t}$	0	9	5	14	0	0	14	0	0	0	3	11	0	0	14
	$A_{6,t}$	0	7	6	6	2	5	3	6	4	9	4	0	0	5	8

	Table 5.6 Continued															
	$A_{6,t}$	5	0	7	0	0	12	4	8	0	0	12	0	1	11	0
	$A_{6,t}$	12	0	0	0	12	0	0	0	12	7	3	2	6	0	6
	$A_{6,t}$	9	2	0	3	8	0	0	2	9	0	0	11	11	0	0
	$A_{6,t}$	0	5	7	4	0	8	6	4	1	11	0	1	4	6	2
5	$\delta_{5,t}$	1	1	0	0	1	1	0	1	1	0	1	1	0	1	1
	$A_{6,t}$	8	0	0	0	0	8	0	0	8	0	4	4	0	0	8
	$A_{6,t}$	2	6	0	0	1	7	0	7	1	0	4	4	0	8	0
	$A_{6,t}$	0	6	0	0	6	0	0	6	0	0	1	6	0	0	6
	$A_{6,t}$	4	0	0	0	4	0	0	0	4	0	4	0	0	4	0

5.7. Dominance of Heuristics

For all 10 problems the decision or solution sets ($\delta_{k,t}$ and $A_{i,t}$) were applied and run on the ASF simulation model with replication set as per table 5.5. The results are shown in table 5.7 which documents both the mean and half-width values for each problem across all five heuristics. As expected the performance function Ω gives different results for the different heuristics, with some decision solutions clearly outperforming the others. Note we consider only the variable cost portion of Ω. For some problems the Ω – half width is relatively small (#7, #9), while for others it is relatively large (#2, #4).

Table 5.7 Simulation Results for Function Ω Variable Costs by Heuristic

Prob #	Heuristic #1		Heuristic #2		Heuristic #3		Heuristic #4		Heuristic #5	
	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width
1	1290	48.1	1020	41.4	1190	46.3	1330	45.3	1090	44.2
2	2136	84.1	1836	68.8	1631	64.3	1636	68.4	2736	106.1
3	1002	34.8	992	35.6	1122	39.8	902	35.2	1242	47.5
4	1592	60.5	992	39.6	1492	61.6	992	37.8	1892	70.3
5	1408	51.8	1157	43.7	1208	51.6	1105	40.6	2308	81
6	1104	36.8	1014	39.3	934	33.5	964	37	1214	38.1
7	950	28.6	910	34.1	920	31	820	25.7	910	34.1
8	1156	42.1	1046	38.2	1516	49.2	1096	38.2	1506	49.2
9	872	29.3	882	30.1	882	30.1	882	30.1	962	36.5
10	1031	36.6	1058	38	1036	37.2	1058	38	1392	45.3

To determine which decision solutions are the best for each problem, the authors conduct a paired t-test for each pair of heuristics in each problem. Based on the Ω-Mean values the rank of each solution is determined, this is shown in table 5.8. Defining the hypothesis test for the first two ranked solutions as follows:

H0: The performance of Rank-1 and Rank-2 heuristics is the same

H1: The performance of Rank-1 and Rank-2 heuristics is not the same

If the null hypothesis is accepted then both decision solutions are included in the optimal solution set. The process is then repeated with the third and fourth ranked solutions. The test results are shown in table 5.8 below.

Table 5.8 Paired t-Test Comparison of Heuristic

	Heuristic Performance #							
Prob #	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**	**Rank 1/2 Ω - Mean**	**Rank 1/2 95% CI Half-Width**	**Heuristics ∈ Optimal Decision Set**
1	2	5	3	1	4	-64.4	57.9	2
2	3	4	2	1	5	-22.8	89.5	3,4
3	4	2	1	3	5	-87.4	48.9	4
4	2	4	3	1	5	-49.3	54.7	2,4
5	4	2	3	1	5	-52.6	58.4	2,3,4
6	3	4	2	1	5	-26.8	44.6	3,4
7	4	2	5	3	1	-90	41.1	4
8	2	4	1	3	5	-47.4	50.6	2,4
9	1	2	3	4	5	-7.76	36.3	1,2,3,4
10	1	3	2	4	5	-26.9	45.9	1,2,3,4

According to the results shown in table 5.8, the authors have the following conclusions:

- Heuristics #2, #3 and #4 show better results than 1 and 5 and are statistically dominant across the set of benchmark problems.
- The asymmetrical load balancing strategy is clearly effective in improving the ASF operation performance.
- Heuristic #4 is the best performing and is in the optimal set for 9 of the 10 problems. Indicating that surgery activity time balance is a significant factor in ASF performance.
- Heuristic #2 also performs well and has an Ω differential ranging from 0% to 13% with an average disadvantage of 5%.
- In combination Heuristics #2 and #4 are a dominant pair by giving the best solutions for the full set of problems.
- Heuristic #5 is the weakest performing with the Ω differential ranging from 7% to 109% with an average disadvantage of 44%.

5.7.1. Cost Component Analysis of Heuristic Solutions

The three cost components comprising the variable cost in were studied further specifically for the heuristic #1, #2 and #4 solutions. The percentage distribution for each component across the first five test problems is shown below in Figure 5.1 to 5.3.

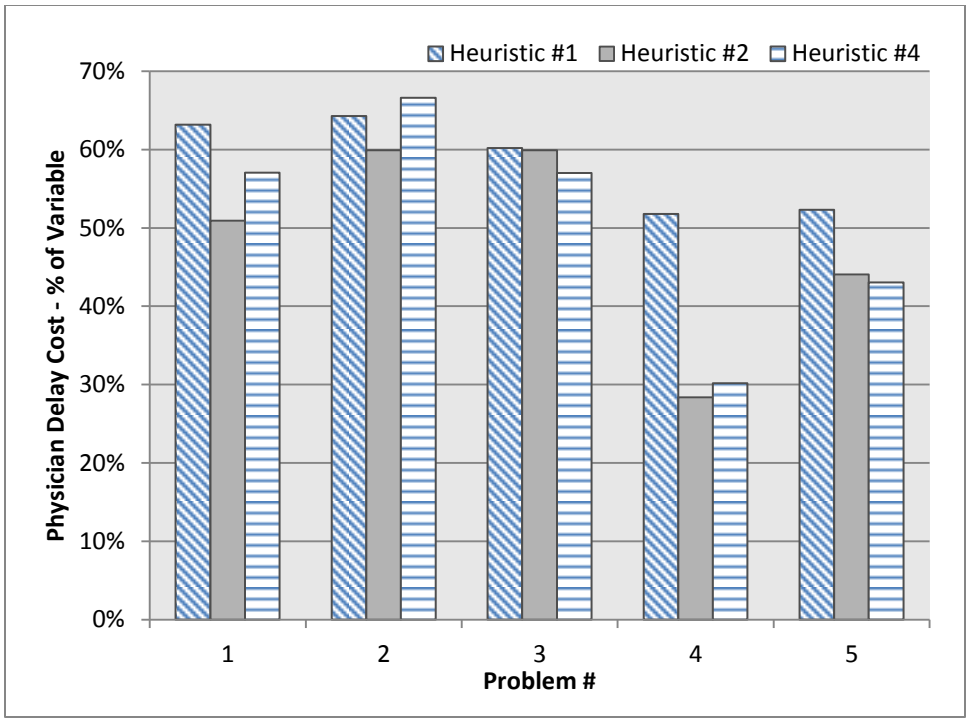


Figure 5.1 Physician Delay as Part of Variable Cost in Ω.

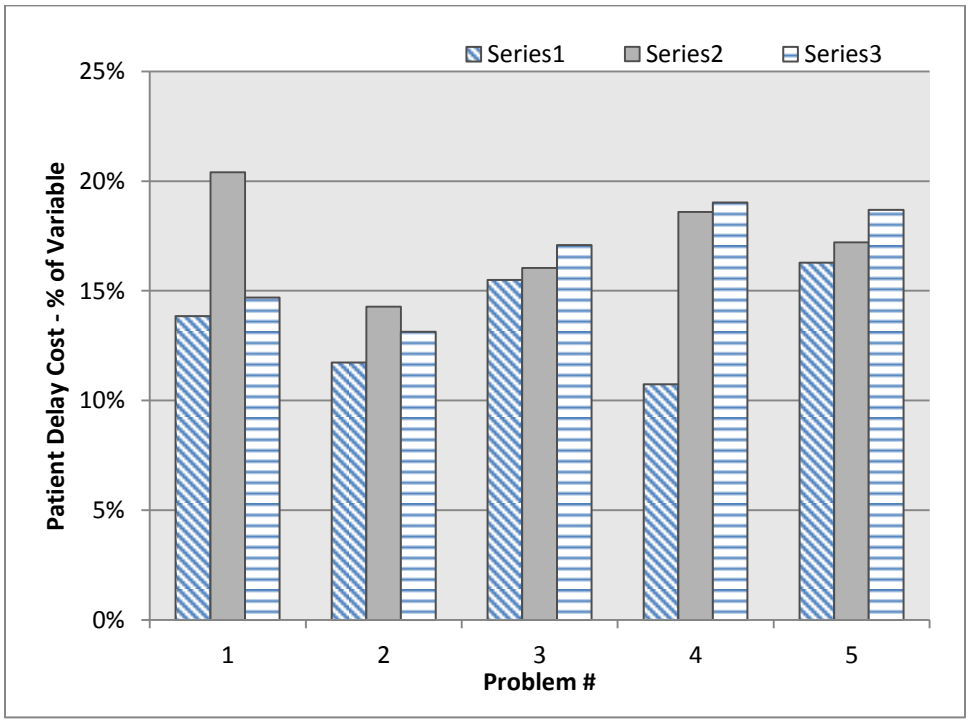


Figure 5.2 Patient Delay as Part of Variable Cost in Ω.

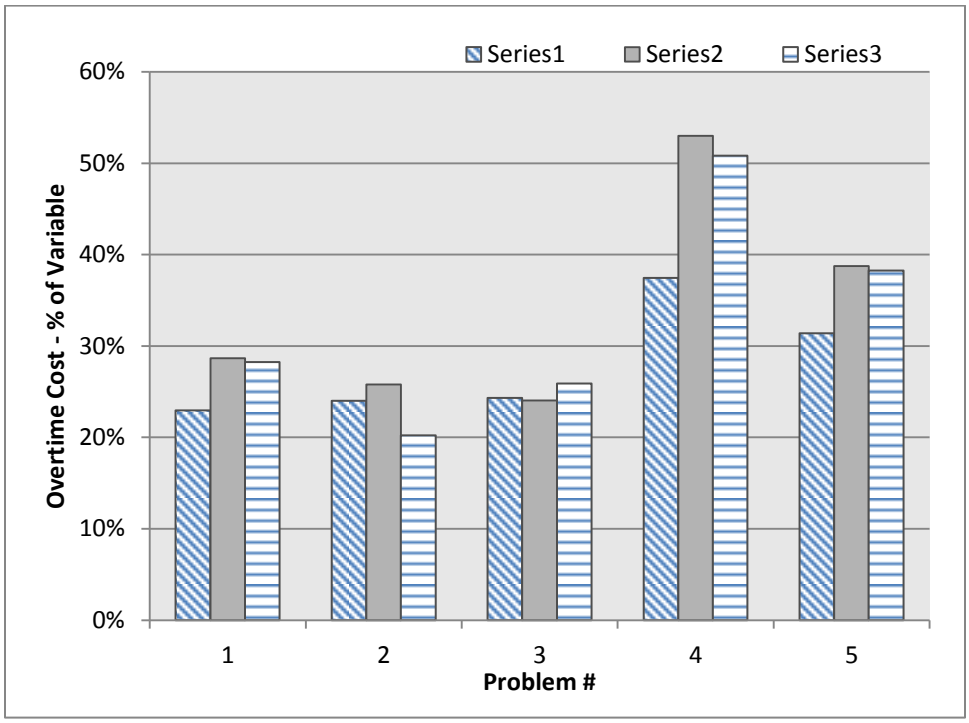


Figure 5.3 Overtime as Part of Variable Cost in Ω.

First the authors evaluate the differences across the five test problems. Problem #2 has on average the highest percentage of physician delay cost. In general problems #1, #2 and #3 have physician delay in the 55-65% range. In contrast problem #4 has a significantly different cost behavior, with overtime being the dominant cost. Both problems #4 and #5 have physician delay less than 50% and the solutions are pressed to resolve the overtime cost. Across the problems patient delay is typically the smallest component and is in the 12-17% range. A key finding of this research is that patient delay is rarely a dominant cost in ASF operations. The authors suspect that this true for many healthcare operations. This is in contrast to the primary focus the authors see on reducing patient waiting time in the healthcare systems research.

Next the authors evaluate the solution differences between heuristics. This is most pronounced in heuristic #1 which tends to have a higher physician delay compared to #2 and #3. Clearly the asymmetrical resource loading strategy is favorable to the physician delay component. This difference is greatest in problem #4 where there is a 20% difference between the solutions. Comparing heuristics #2 and #4, problem #1 is most interesting. By balancing the surgery times the authors see that heuristic #4 raises physician delay by 6% but gives a 7% reduction in the patient delay. This behavior is not consistent though across all the problems. The results confirm that the heuristics each behave uniquely and do have different strategies across the problems.

5.8. Deriving a Lower Bound to Ω

A key issue in the performance analysis of heuristics is to have a good estimate of the lower bound solution, and this allows us to gauge the true quality of the solutions. Where an exact solution is available then that becomes the lower bound. Following the experience and with the ASF model and its operations, the authors find the following two methods can give us ways to get lower bound for doctor's delay, patient waiting time and resource overtime cost and both of these methods will be applied to the original best results:

- ***Lower Bound 1 (LB-1):*** By increasing the staffing resource levels M_j by 20% for al *j*. Since the three variable cost objectives are inverse to the fixed staffing cost, by relaxing the staffing constraint the variable cost will drop. This revised variable cost then serves as the lower bound.

- ***Lower Bound 2 (LB-2):*** The model study shows that patient arrival times are a key factor in performance. By initiating patient arrivals into the system 15 minutes earlier than the original setting this relaxes then patient arrival constraint. As a result the variable costs of overtime and physician delay will go down, though the patient delay cost will go up. The overall lower variable cost then serves as the lower bound.

Let Ω* represent the best solution generated from the heuristics. The associated physician assignment solution ($\delta_{k,t}$ and $A_{i,t}$) is then rerun with the relaxed constraints listed above. There are therefore three solutions to each problem Ω*, LB-1and LB-2. For each solution the three variable costs components are tracked: Physician Delay, Patient Delay and Overtime. The overall lower bound LB* is then give by the sum of the minimum of each component for the three solution. That is:

$$LB^* = Min\{Physician\ Delay: \Omega^*, LB1, LB2)\} + Min\{Patient\ Delay: \Omega^*, LB1, LB2)\} + Min\{Overtime: \Omega^*, LB1, LB2)\}$$

Table 5.9 below provides LB* for the test problems and compares it against the best heuristic solution. The LB* gap ranges from 3.26% from 7th problem to 27.04% from the 10th problem. The highest three gaps are from problems #3, #8 and #10 which are all above 20% relative to LB*.

Table 5.9 Lower Bound Gap Analysis

Prob #	Best Heuristic	Solution	Physician Delay ($)	Patient Delay ($)	Overtime ($)	Total Cost ($)	Gap
1	#2	Ω*	566	238	216	1020	**18.69%**
		LB*	431	204	194	829	
2	#3	Ω*	980	289	362	1631	**7.78%**
		LB*	894	279	331	1504	
3	#4	Ω*	509	154	239	902	**20.43%**
		LB*	366	151	201	718	
4	#2	Ω*	178	245	569	992	**14.09%**
		LB*	120	239	493	852	
5	#4	Ω*	447	221	437	1105	**14.12%**
		LB*	355	215	379	949	
6	#3	Ω*	531	207	196	934	**11.87%**
		LB*	464	179	181	823	
7	#2	Ω*	315	265	240	820	**3.26%**
		LB*	315	246	232	793	
8	#2	Ω*	636	203	207	1046	**22.39%**
		LB*	465	177	169	812	
9	#2	Ω*	605	197	80	882	**14.96%**
		LB*	514	169	67	750	
10	#2	Ω*	668	196	172	1036	**27.04%**
		LB*	451	160	145	756	

5.9 Results Data Analysis

To explain the reason why the three problems have comparably higher gaps to others, the authors investigate the nature of 10 problems by tracking several factors from the

problems' themselves. From the fourth (total pre-operation time in minutes over total patient number) and sixth factor (total pre-operation time in minutes over total staffing numbers) the authors listed in the left column, p3, p8 and p10 gives higher numbers than any problems else. Because of the heavier pre-operation load than other problems, these three cases may decrease more when the authors arrange more staffing members in the system and make patients come earlier.

Table 5.10 Lower Bound Gap Analysis

Category\Problems	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10
1 Total patient out	144	274	116	350	286	78	140	133	36	74
2 Total pre opt time	6672	12407.70	5359.48	16767.70	13468.69	3578.83	6490.39	6528.80	1610.97	3895.25
3 Total opt time	4899	9855.28	3998.88	12567.24	10340.46	2763.72	4781.39	4239.40	1212.84	2588.76
4 Pre-opt time/patient	46.33	45.28	46.20	47.91	47.09	45.88	46.36	49.09	44.75	52.64
5 opt time/ patient	34.02	35.97	34.47	35.91	36.16	35.43	34.15	31.88	33.69	34.98
6 pre opt time/staff	202.19	206.79	223.31	212.25	207.21	210.52	216.35	233.17	179.00	229.13
7 opt time/staff	148.47	164.25	166.62	159.08	159.08	162.57	159.38	151.41	134.76	152.28
8 total staff number	33.00	60.00	24.00	79.00	65.00	17.00	30.00	28.00	9.00	17.00
9 blocks taken	25.00	45.00	21.00	63.00	54.00	15.00	24.00	21.00	6.00	13.00
10 opt time /blocks taken	0.82	0.91	0.79	0.83	0.80	0.77	0.83	0.84	0.84	0.83

1Total patient out: $\sum_{i=30} \alpha_i$

2Total pre opt time: $\sum_{i=30} \alpha_i * (\mu_{i,1} + Z_{0.75} * \sigma_{i,1})$

3Total opt time: $\sum_{i=30} \alpha_i * (\mu_{i,2} + Z_{0.75} * \sigma_{i,2})$

4Total pre opt time/patient: $\sum_{i=30} \alpha_i * (\mu_{i,1} + Z_{0.75} * \sigma_{i,1}) / \sum_{i=30} \alpha_i$

5Total opt time/ patient: $\sum_{i=30} \alpha_i * (\mu_{i,2} + Z_{0.75} * \sigma_{i,2}) / \sum_{i=30} \alpha_i$

6 Total pre opt time/staff: $\sum_{i=30} \alpha_i * (\mu_{i,1} + Z_{0.75} * \sigma_{i,1}) / \sum_{t}^{t=3} \sum_{j}^{j=3} M_{j,t}$

7Total opt time/staff: $\sum_{i=30} \alpha_i * (\mu_{i,2} + Z_{0.75} * \sigma_{i,2}) / \sum_{t}^{t=3} \sum_{j}^{j=3} M_{j,t}$

8 Total staff number: $\sum_{t}^{t=3} \sum_{j}^{j=3} M_{j,t}$

9 Total blocks taken: $\sum_{t}^{t=3} \sum_{k}^{k=6} \delta_{k,t} * N_k$

10 Total opt time / Total blocks taken: $\sum_{i=30} \alpha_i * (\mu_{i,2} + Z_{0.75} * \sigma_{i,2}) / \sum_{t}^{t=3} \sum_{k}^{k=6} \delta_{k,t} * N_k$

CHAPTER 6

SCHEDULING ARRIVAL TIMES OF INDIVIDUAL PATIENTS

The physician assignment problem discussed in chapter 5 included the derivation of the arrival volume of all patients $A_{i,t}$ in a time block. The common approach is to convert the sum of all for a physician group k into a series of equal interval arrivals. A base strategy is to set the interval in the 20-30 minute range. In this chapter the authors expand further on by (i) identifying the patient arrival sequences (ii) dynamic setting of the inter arrival time between every pair of patients and (ii) prescribing the arrival time of the first patient in each time block. In this chapter, new updates to the simulation model were made to adapt to new specific appointment time for a specific patient. Considering the surgery process as flow shop problem, some classic flow shop heuristic rules were adapted to the problem here.

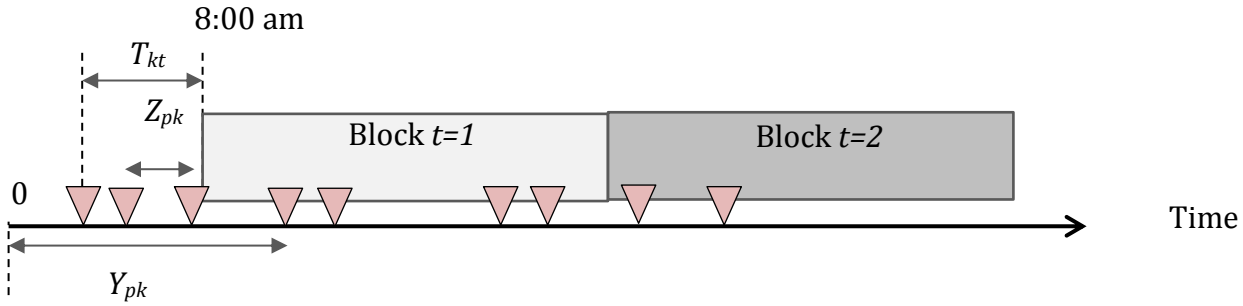
6.1. Defining the Patient Arrival Time Scheduling Problem

The results from the simulations experiments in chapter 4 and 5 have demonstrated the importance of patient arrival rates and mix to the overall objective function. The results have also shown that contrary to common beliefs patient delay costs are not dominant. It is unlikely that the authors will see close to zero patient waiting times in almost any healthcare service facility. This is primarily due to the order of magnitude difference between physician delay costs and patient delay costs. There is much interest know in

scheduling patient arrivals more specifically, that is each patient is given a specific arrival time which minimize their project delay time. In the extreme case these arrival are scheduled in real time, patient are given an arrival window and then as the effect of the system variance becomes more apparent the time is updated in real time. The patient arrival time scheduling problem therefore involves prescribing the specific arrival time of each patient in the system, which then functions as the expected arrival time of that patient. The problem decisions are then as follows:

p	Sequence number of patient arriving for each physician group
$T_{k,t}$	The early arrival buffer for the first patient in block t for group k
$Z_{p.k}$	The expected inter-arrival time between patient p-1 and p for group k
$Y_{p.k}$	The expected arrival time of patient p for group k

The schematic relationship of these decision variables is in shown in Figure 6.1, including an example decision table. Note that $Z_{p.k}$ is a dependent decision, since $Z_{p.k}$ = $Y_{p.k}$ - $Y_{p-1.k}$. Depending on the decision strategy one or the other is determined first.



Sequence p	1	2	3	4	5	6	7	8	9
Type i	7	7	7	4	4	3	3	9	9
$Y_{p.k}$ *hours*	-0.75	-0.42	-0.15	0.80	1.08	3.25	3.60	4.50	5.40

Figure 6.1 Patient arrival time decisions and relationships.

The ASF patient arrival time scheduling problem is then described as determining the sequence p, $T_{k,t}$, and $Y_{p.k}$ for each physician group *k*, such that the expected value of Ω is minimized. Further the authors assume the staffing level has already been fixed, that is $M_{j,t}$ is predetermined. The ASF operational objective has been previously defined as follows:

$$Minimize: \quad \Omega = \sum_{j}\sum_{t}\left(4\varphi_{j,R}M_{j,t}\right) + \sum_{j}\left(\varphi_{j,O}O_{j}\right) + \phi_{P}T_{P} + \phi_{D}T_{D}$$

Since $M_{j,t}$ is fixed then the first term in Ω is a constant for a given problem. The authors present here several heuristics based on classical flow shop scheduling rules to solve this problem.

6.2. Review of Flowshop Scheduling and Sequencing

There are different ways to classify scheduling and sequencing problems, commonly based on the quantity of machines, quantity of jobs, or the objective function etc. In the multiple machines' category, according to the characteristics of jobs and operation orders, job-shop and flow-shop problem have been classified as two different types.

Job-shop problem is a number of jobs have to be done and every job consists of using a number of machines for a certain amount of time. Flow shop scheduling problem is with a work shop or group shop in which the flow control shall enable an appropriate sequencing for each job and for processing on a set of machines or with other resources

1,2,...,m in compliance with given processing orders. The job shop process differs from flow shop process in that the flow of work is not unidirectional in job shop, hence it is one of the complex scheduling problems. One interesting example of job-shop problem is the traveling salesman problem. The travelling salesman problem (TSP) asks the following question: Given a list of cities and the distances between each pair of cities, what is the shortest possible route that visits each city exactly once and returns to the origin city? It is an NP-hard problem in combinatorial optimization, important in operations research and theoretical computer science. Then the job-shop problem is clearly also NP-hard, since the TSP is special case of the JSP with m = 1 (the salesman is the machine and the cities are the jobs.)

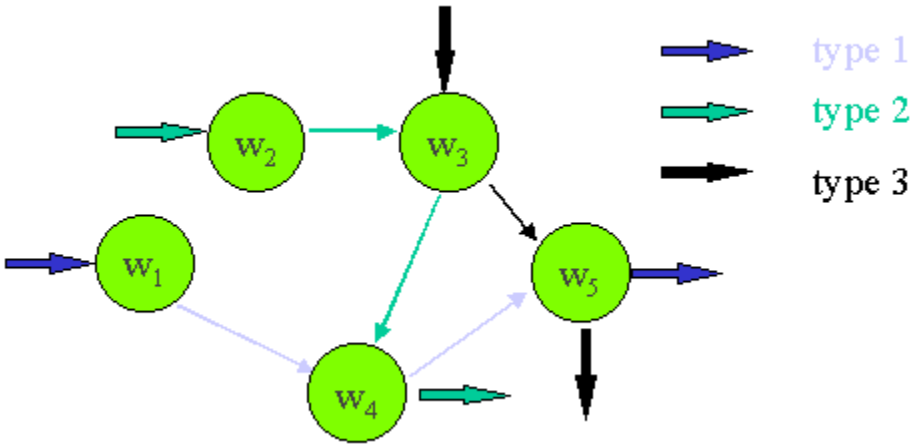


Figure 6.2 job-shop

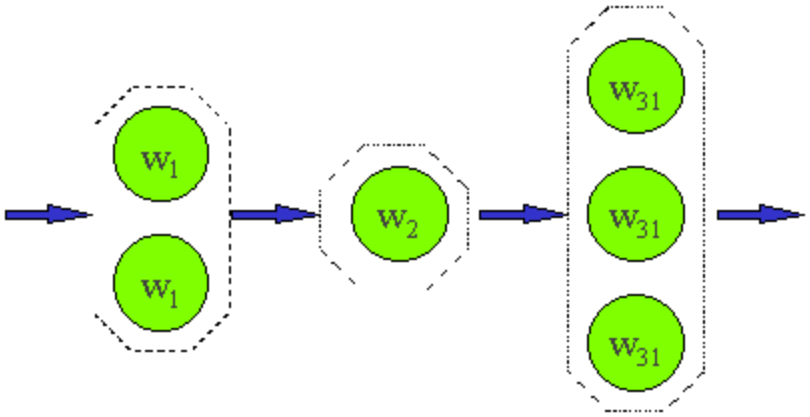


Figure 6.3 Flexible flow-shops.

Typical specification of scheduling problems involves first specifying the problem in the format:

N/M/A/B

N: 1, 2, or N (number of jobs)

M: 1, 2, or M (number of machines)

A: the job flow pattern (discussed later in these notes), and

B: the performance measure (e.g. average flow time)

NOTE: The performance measure is always stated in terms of a measure that needs to be MINIMIZED, so the B-field only shows the criterion, not the optimization condition.

Usually, the authors are interested in N-jobs problems, and therefore the authors may specify only the last three of these in the problem, that is, M/A/B specification. Some people call this the α/β/γ specification)

Notations:

α machine environment

β processing characteristics/constraints

γ objective functions

Table 6.1 α -field notations

1	Single machine
Pm	m identical machines in parallel; job-j can be processed by any one of them.
Qm	m machines in parallel with different processing time; job-j can go to any one.
Rm	m unrelated machines in parallel; each job can go to a particular one (or one of a subset) of these m.
Fm	Flow shop: m machines in series. Each job goes to each machine. All jobs have same routing. Most common schedule is a permutation schedule (FIFO at each machine)
FFs	Flexible flow shop. S-stages in series, each stage has 1, 2, or more machines, and each job may be assigned to exactly one machine in each stage.
Om	Open shops. Each job must visit each of m machines; each job has unique route; If jobs can visit same machine more than once, then Om|recrc|
Jm	m machine job shop. Each job can visit one or more machine, and has its own route.

Table 6.2 β -field notations

rj	Release dates are specified
sjk	Sequence dependent setup times: setup time between job j and job k depends on machine.
prmp	Preemptions. A job being processed can be interrupted by another (remaining operations must be completed at a later time).
prec	Precendeces are specified, as a set A of pairs (j,k) if task j is an immediate predecessor of task k.
brkdwn	Breakdown: machines are not available continuously; in deterministic scenarios, scheduled maintenance shut downs can be modeled thus.
Mj	Machine eligibility restrictions; Mj is th set of machines that can do task j.
prmu	Permutation schedule: in flow shops -- each machine processes tasks in FIFO.
block	Blocking. If buffer before machine j is full, then upstream machine cannot release task.
recrc	Recirculation: job can visit same machine more than one time.
no-wait	Streaming: jobs cannot wait between one process and next.

NOTE: while any given problem specification will contain only one entry in the a-field, it may contain several comma-separated values in the b-field.

γ -field notations:

Cj : Completion time

Lj : Lateness (Cj - dj)

Tj : Tardiness, (= max{ 0, Lj}).

Uj : Unit penalty (= 1, if Tj is non-zero, 0 otherwise)

Fj: The flow time of a job, the time the job is in the system, Fj=Cj-rj

Fmax: Maximum flow time in the system

Cmax : Makespan (= completion time of the last task of the last job to be completed)

Lmax : Maximum lateness.

ΣwjCj= total weighted completion time. Weight of each job indicates its relative importance.

ΣwjTj = total weighted tardiness.

ΣwjUj= total weighted unit penalty.

Each of the above are objectives that must be minimized. Typically, a shop manager will select his/her favorite from among these objectives to schedule their factory. Most of these are self-explanatory, except for the last one. There, the authors see use, instead of a linear importance of the completion time, a non-linear function, which depending upon the rate, r, and makes jobs with closer completion times relatively more important than others.

6.3. Heuristic Algorithms for Three Machine Problem

Considering the three stage process for the whole surgery (preoperative, surgery and postoperative) to be three machines flexible flow-shop problem with release time rj,

permutation schedule, set up time and block constraints, targeting at minimal total weighted tardiness time. However, minimizing the total tardiness on one machine is NP-hard (Leung, 1990), different heuristic algorithms have been promoted to achieve both less calculation complexity and enough good solutions.

However, the ASF surgery scheduling problem is unlike any of the classical minimize of the total tardiness problem, and the multi objectives are not just equal to the total tardiness. Therefore, other famous three machine flow shop heuristics have been reviewed and tested. Three heuristic algorithms to solve the n/3/P/Makespan problem have been generally used and also been tested on the simulation model, and they are Cambell-Dudek-Smith's Rule, Palmer's Rule and Critical Path Method. Though not in optimizing the total weighted tardiness, heuristics based on these rules gave good results from the ASF experiments, not only in general total cost but also in separate single objectives. Seven heuristic are generated and five of their results are shown in the following sections. The general ideas for these rules are:

1. Find the first arrival time
2. Calculate the time intervals
3. Decide the patient group sequence
4. Assign overall patient

Let p = 1 to U_k, where $U_k = \sum \alpha_i$ for group k, that is total patient for group k

Model variables are introduced as follows:

S_t is the window start time for each window t

Z_{pk} is the inter arrival time between patient p and p-1 (hour)

T_{kt} is the early time for first patient (hour)

Y_{pk} is the planned arrival time of patient p for group k (hour)

SPT_i is the total processing time for patient type i

ζi is the ratio of Palmer's Rule

6.3.1. Heuristic #1- Based on Cambell-Dudek-Smith CDS

In the chapter 5, A_{it} is used to stand for patient arrival rate, however here the specific arrival time Y_{pk} (like 7.5 hour) will be specified.

Step – 1: Calculate the first arrival time for the physician group k based on $\delta_{k,t}$ where $\delta_{k,t}$ is the same as the dummy problem. In the daily three of the four hour window: 8:00-12:00, 12:00-16:00, and 16:00-18:00, the patients are required to arrive 45 minutes (0.75 hour) earlier to the system to finish the documentation.

When the first $\delta_{k,t} > 0$ for physician group k appears on window t,

$$T_{kt} = S_t - 0.75$$

If t=1, T_{k1}=8-0.75=7.25 (hour)

If t=2, T_{k2}=12-0.75=11.25(hour)

If t=3, T_{k2}=16-0.75=15.25(hour)

Step – 2: Calculate the inter arrival time Z_{pk}, because the authors have 4 hour window for each physician group, but all the patients are allowed only to arrival within 3 hours to reduce post-operative delay.

$$Z_{pk} = {3 \times \sum_k \delta_{k,t}}\Big/{U_K}$$

Step – 3: Patient group sequence decision. The CDS algorithm uses Johnson's Rule in a heuristic fashion and creates several schedules from which a "best" schedule is chosen. The algorithm corresponds to a multistage use of Johnson's Rule applied to a two-machine pseudo-problem derived from the original. The Johnson's rule will be simply introduced for CDS explanation.

Johnson's rule is as follows:

1. List the jobs and their times at each work center.
2. Select the job with the shortest activity time. If that activity time is for the first work center, then schedule the job first. If that activity time is for the second work center then schedule the job last. Break ties arbitrarily.
3. Eliminate the shortest job from further consideration.
4. Repeat steps 2 and 3, working towards the center of the job schedule until all jobs have been scheduled.
5. Given significant idle time at the second work center (from waiting for the job to be finished at the first work center), job splitting may be used.

Each of five jobs needs to go through work center A and B. Find the optimum sequence of jobs using Johnson's rule. To solve three machines problem, CDS follows Johnson's rule by either not consider the middle processing time or by adding the first

and second and the second and the third processing time, then pick up the smaller makespan by comparing these two methods. Though CDS targets at minimize makespan, same ideas and format could also be copied to the ASF problem. The following is an example for CDS's rule, it applies Johnson's rule without considering the operative time.

Example.1

Table 6.3 physician group v's patient types

Patient Type i	75% process time (min)			
	pre	op	post	total
9	28	31	65	124
10	51	31	65	147
19	51	60	130	241
20	75	60	130	265

1. Find the smallest pre or post processing time within one physician group, which I highlighted here (28). Because it appears in the first preoperative process, so patient type 9 will be put to the earliest. (If it appears in the post process, the patient type 9 will be put to the end.)

10	51	31	65	147
19	51	60	130	241
20	75	60	130	265

2. Remove patient type9; find the smallest number among patient 10, 19 and 20. If there are two numbers are equal, for this problem, just randomly pick one to process first. Then patient type 10 will be scheduled next.

19	51	60	130	241
20	75	60	130	265

1. Then same way, patient 19 will be followed by patient 10
2. So the final sequence is 9,10,19,20.

Step –5: Decide arrival time for patient p for physician group k

$$Y_{pk} = T_{kt}\ when\ p = 1$$

$$Y_{pk} = T_{kt} + Z_{pk}\ when\ p > 1$$

$$if\ Y_{pk} \geq T_{k,t+1} - 1hour, let\ Y_{p,k} = T_{k,t+1}$$

Step – 6: Overall patient sequence for physician group k:

According to the step 3 on group sequence scheduling results: 9, 10, 19, 20, all type 9 patients got the higher priority to be scheduled first. The $\delta_{k,t}$ is given in the Table 6.4

Table 6.4 Given $\delta_{k,t}$

k\t	t=1	t=2	t=3
1	1	1	0
2	1	1	0
3	0	0	1
4	1	1	1
5	0	1	1
6	0	0	0

T_{11}=8-0.75=7.25 (hour)

And T_{12}=12-0.75=11.25 (hour)

If $Y_{pk} > (T_{12}-1) = 10.25$, let Y_{pk}=11.25

Overall performance for physician group v is shown below in Table 6.5

Table 6.5 Overall Performance Results

patient type i	physician type k	arrival time (min)	arrival time Y_{pk} (hour)
9	**v**	**435**	**7.25**
9	v	453.9474	7.57
9	v	472.8947	7.88
9	v	491.8421	8.20
10	v	510.7895	8.51
10	v	529.7368	8.83
10	v	548.6842	9.14
10	v	567.6316	9.46
10	v	586.5789	9.78
19	v	605.5263	10.09
19	**v**	**675**	**11.25**
19	v	693.9474	11.57
19	v	712.8947	11.88
19	v	731.8421	12.20
20	v	750.7895	12.51
20	v	769.7368	12.83
20	v	788.6842	13.14
20	v	807.6316	13.46
20	v	826.5789	13.78

6.3.2. Heuristic #2-Cycle CDS

Instead of having patient group sequencing rule, the cycle CDS made the single patient sequence the same as CDS rule, but the overall sequence is the repeat of the single sequence result.

Step – 1-5: Same as Heuristic #1

Step – 6: Overall patient sequence for all physician groups.

Table 6.6 Overall Performance Results for Cycle CDS

patient type i	physician type k	arrival time (min)	arrival time Y_{pk} (hour)
9	**v**	**435**	**7.25**
10	v	455	7.58
19	v	475	7.92
20	v	495	8.25
9	v	515	8.58
10	v	535	8.92
19	v	555	9.25
20	v	575	9.58
9	v	595	9.92
10	**v**	**675**	**11.25**
19	v	695	11.58
20	v	715	11.92
9	v	735	12.25
10	v	755	12.58
19	v	775	12.92
20	v	795	13.25
9	v	815	13.58
10	v	835	13.92

6.3.3. Heuristic #3-Batch CDS

Based on the number of physicians in the group, ignoring about other limited resources like the operative beds, the batch of arrival patients are designed to the ASF at the same time.

Step – 1-4: Same as Heuristic #1

Step –5: Decide arrival time for patient p for physician group k

$$Y_{pk} = T_{kt}\ when\ p = 1$$

$$Y_{pk} = T_{kt} + Z_{pk}\ when\ p = 2g + 1, g \in integer$$

$$if\ Y_{pk} \geq T_{k,t+1} - 1hour, let\ Y_{p,k} = T_{k,t+1}$$

Step – 6: Overall patient sequence for all physician groups:

Table 6.7 Overall Performance Results for Batch CDS

patient type i	physician type k	Time intervals (min)	arrival time Z_{pk} (min)	arrival time Y_{pk} (hour)
9	**v**	**0**	**435**	**7.25**
9	**v**	**0**	**435**	**7.25**
9	v	40	475	7.92
9	v	0	475	7.92
9	v	40	515	8.58
10	v	0	515	8.58
10	v	40	555	9.25
10	v	0	555	9.25
10	v	40	595	9.92
10	v	0	595	9.92
19	**v**	**0**	**675**	**11.25**
19	**v**	**0**	**675**	**11.25**
19	v	40	715	11.92
19	v	0	715	11.92
20	v	40	755	12.58
20	v	0	755	12.58
20	v	40	795	13.25
20	v	0	795	13.25

6.3.4. Heuristic #4-Modified CDS

Other steps are the same as heuristic #1 except step 3

Step – 3: Patient group sequence decision

If there are two types of patients have the same preoperative time which for example is shown below as patient type 10 and patient type 19. However, the postoperative time is the different for them. Unlike the first heuristic, some considerations will be made based on the different postoperative time even the preoperative time are the same for them.

Because the main ideal of the Johnson's rule is: made the shorter processing time's part go either front or to the end. When under the same shortest preoperative time, but patient type 10's post-operative time is smaller than its of patient type 19's, so patient type 10 should be put to later place and patient type 19 should be put to the second place.

10	51	31	65	147
19	51	60	130	241
20	75	60	130	265

So the final patient group sequence for this heuristic is 9,19,10,20

Step – 6: Overall patient sequence for physician groups.

Table 6.8 Overall Performance Results for Modified CDS

patient type i	physician type k	arrival time (min)	arrival time Y_{pk} (hour)
9	v	435	7.25
9	v	455	7.58
9	v	475	7.92
9	v	495	8.25
9	v	515	8.58
19	v	535	8.92
19	v	555	9.25
19	v	575	9.58
19	v	595	9.92
10	v	675	11.25
10	v	695	11.58
10	v	715	11.92
10	v	735	12.25
10	v	755	12.58
20	v	775	12.92
20	v	795	13.25
20	v	815	13.58
20	v	835	13.92

6.3.5. Heuristic #5-Total Time SPT

Other steps are the same as heuristic #1 except step 3

Step – 3: Patient group sequence decision: give the shortest total processing time the highest priority to be scheduled.

$$\mathrm{SPT}_i = \sum_n (\mu_{i,n} + Z_{0.75} \times \sigma_{i,n})$$

Based on the total processing time SPT_i which is shown in the last column (total), smaller total processing time's patient type will be put to the earlier place.

Example 2

Patient Type i	75% process time (min)			
	pre	op	post	SPT_i
9	28	31	65	**124**
10	51	31	65	**147**
19	51	60	130	**241**
20	75	60	130	**265**

Because $SPT_9 < SPT_{10} < SPT_{19} < SPT_{20}$, the final group sequence for this heuristic is 9,10,19,20

Step – 6: Overall patient sequence for physician group v:

Table 6.9 Overall Performance Results for Total Time SPT

patient type i	physician type k	arrival time (min)	arrival time Y_{pk} (hour)
19	**v**	**435**	**7.25**
19	v	455	7.58
19	v	475	7.92
19	v	495	8.25
20	v	515	8.58
20	v	535	8.92
20	v	555	9.25
20	v	575	9.58
9	v	595	9.92
9	**v**	**675**	**11.25**
9	v	695	11.58
9	v	715	11.92
9	v	735	12.25
10	v	755	12.58
10	v	775	12.92
10	v	795	13.25
10	v	815	13.58
10	v	835	13.92

6.3.6. Heuristic #6- Palmer

Other steps are the same as heuristic #1 except step 3

Step – 3: Patient group sequence decision

Palmer's Rule is based on the ratio ζi for different processing time, and sequences the highest priority job first. When for three machine problems, the equation is shown as follows:

$$\boldsymbol{\zeta_i = \mu_{i,3} - \mu_{i,1}}$$

Example.3

Table 6.10 Physician Group V's patient types

Patient Type i	75% process time (min)				
	pre	op	post	ζi	total
9	28	31	65	**37**	124
10	51	31	65	**14**	147
19	51	60	130	**79**	241
20	75	60	130	**55**	265

Because $\zeta_{19} > \zeta_{20} > \zeta_9 > \zeta_{10}$, the final sequence is 19, 20, 9, and 10.

Step – 6: Overall patient sequence for physician group v:

Table 6.11 Overall Performance Results for Palmer

patient type i	physician type k	arrival time (min)	arrival time Y_{pk} (hour)
19	**v**	**435**	**7.25**
19	v	455	7.58
19	v	475	7.92
19	v	495	8.25
20	v	515	8.58
20	v	535	8.92
20	v	555	9.25
20	v	575	9.58
9	v	595	9.92
9	**v**	**675**	**11.25**
9	v	695	11.58
9	v	715	11.92
9	v	735	12.25
10	v	755	12.58
10	v	775	12.92
10	v	795	13.25
10	v	815	13.58
10	v	835	13.92

6.3.7. Heuristic #7-Critical Path Method

Other steps are the same as heuristic #1 except step 3

Step – 3: Patient group sequence decision

- Find the highest total processing time, and consider the according patient type as the critical type (C)
- for the rest parts, if the first processing time is less than the last processing time, sequence the shortest first processing time first(S1).
- Otherwise, sequence the highest last processing time first (S2).
- The whole sequence has been composed by (S1, C, S2)

Example 4

Table 6.12 Physician Group V's patient types

Patient Type i	75% process time (min)			
	pre	op	post	total
9	28	31	65	124
10	51	31	65	147
19	51	60	130	241
20	75	60	130	265

1. Find the highest total processing time which is patient type 20, and that is the critical patient type.
2. For the rest of the patient types, if $\mu_{i,1}<\mu_{i,3}$, which includes patient type 9, 10 and 19, make an order based on increasing $\mu_{i,1}$, which is 9, 10, 19.
3. If $\mu_{i,1}>\mu_{i,3}$, by not increasing $\mu_{i,3}$, make the order.
4. If there are two numbers are the same, test it and choose the better one.

The final sequence for this one is 9,19,10,20

Step – 6: Overall patient sequence for physician group

Table 6.13 Overall Performance Results for Critical Path Method

patient type	physician type	arrival time (min)	arrival time (hour)
9	v	435	7.25
9	v	455	7.58
9	v	475	7.92
9	v	495	8.25
9	v	515	8.58
19	v	535	8.92
19	v	555	9.25
19	v	575	9.58
19	v	595	9.92
10	v	675	11.25
10	v	695	11.58
10	v	715	11.92
10	v	735	12.25
10	v	755	12.58
20	v	775	12.92
20	v	795	13.25
20	v	815	13.58
20	v	835	13.92

6.4. Replication Estimate for Experiments

Because of the overall performance, the results analysis is only from Heuristice#1, #4, #5, #6 and#7. In Heuristic #2, because of different α_i, some long processing time patient types are arranged in the last which leads to high overflow cost. In Heuristic #3, it has effective on increasing the usage of physician groups, however, it results to the overcrowded queue for preoperative.

To have more precise experimental results, the initial 100 replications experiments have been applied and the number of estimated replications has been

calculated by the same way which has been introduced in Chapter 5. The following table 6.14cis the calculation steps in details for the first heuristic under 100 initial replications.

Table 6.14 Replication Estimate for Experiments Table

Problem	Total Cost	Fixed Cost	Variable Cost	4%Ofvariable	Half Width	Estimate Replication Time
1	8861.69	7620	1241.69	49.6676	68.94	192.66
2	16981.45	15336	1645.45	65.818	84.6	165.22
3	7572.77	6924	899.24	35.9696	43.51	146.32
4	24794	21996	2798	111.92	119.65	114.29
5	19965	17856	2109	84.36	35.52	17.73
6	6694.74	4872	1822.74	72.9096	89.77	151.60
7	7986.96	6960	1026.96	41.0784	66.7	263.65
8	10565.65	8256	2309.65	92.386	70.87	58.85
9	2993.08	2784	209.08	8.3632	8.39	100.64
10	6584.4	4740	1844.4	73.776	93.03	159.01

The following Table 6.15 is the maximum number of replications among five heuristics, and this has been applied to all the problems with different heuristics to get smaller variance number.

Table 6.15 Replication for Experiments Table

Problem	1	2	3	4	5	6	7	8	9	10
Reps	220	200	250	130	230	380	260	180	130	340

6.5. Total Cost Comparison and Conclusion

Table 6.16 shows the rank of five algorithms for total ten problems. Then paired t test has been applied to test the difference of the original experimental results with certain replication settings from table 6.15. The first step is to test the difference between rank 1 and rank 2 results with the hypothesis test below. If the statistical results show that rank 1

is lower than rank 2, the rank 1 algorithm is proved to give the optimal results among five algorithms. If it fails to reject the H0, the next step is to compare the rank 2 and rank 3 algorithm results, and equal algorithm results would be shown in this case. The following table 6.17 displays the final results with mean and half width in details, and highlighted ones are the best.

H0: Rank 2 algorithm is equal to rank 5 algorithm

H1: Rank 2 algorithm is not equal to rank 5 algorithm

Table 6.16 Hypotheses for Ten Problems

	Heuristic Performance							
Prob #	**Rank 1**	**Rank 2**	**Rank 3**	**Rank 4**	**Rank 5**	**Rank 1/2 Ω - Mean**	**Rank 1/2 95% CI Half-Width**	**Heuristics ∈ Optimal Decision Set**
1	2	5	3	1	4	-391	36.7	2
2	3	4	1	5	2	-324	65	3
3	1	3	5	2	4	-10.8	28	1,3
4	3	4	5	2	1	-96.2	77.2	3
5	4	3	5	1	2	-369	40.7	4
6	4	5	2	1	3	-106	59.6	4
7	2	5	4	1	3	-8.86	27.7	2,5
8	4	2	5	3	1	-143	81.6	4
9	3	2	5	1	4	-15.1	4.17	3
10	4	3	1	5	2	-594	61.3	4

Table 6.17 Final Results

Prob #	Heuristic #1		Heuristic #2		Heuristic #3		Heuristic #4		Heuristic #5	
	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width	Ω - Mean	Ω - Half Width
1	1253.73	45.88	691.822	20.93	1208.56	49.49	1326.48	59.15	1082.73	33.33
2	1650.41	58.92	2419.17	94.47	1121.38	36.7	1445.65	53.18	1697.34	74.9
3	649.465	23.1	971.788	29.63	660.307	21.13	2045.28	76.04	755.327	18.44
4	2751.1	102.27	2079.11	78.45	1455.55	52.97	1551.75	57.22	2024.76	80.7
5	2120.16	81.93	2581.36	91.63	1262.8	38.39	893.473	15.71	2107.6	83.46
6	1879.66	49.02	1634.42	45.2	2064.37	55.83	1420.39	53.46	1526.16	41.87
7	1050.85	40.11	813.963	24.65	1100.24	32.76	914.503	31.67	822.813	18.38
8	2613.47	51.67	2017.33	75.42	2307.69	51.67	1874.56	52.53	2077.89	84.04
9	245.15	6.78	204.572	5.16	189.64	4.81	261.466	11.69	223.311	4.99
10	1851.34	51.53	1887	48.82	1770.25	50.75	1176.34	45.31	1866.68	43.35

There is no dominance, SPT and Palmers are equally good. There are some proofs which may show that SPT has effects on patient waiting time cost, and the palmer's rule has reduced the overflow cost. With more accurate arrival time, simulation variance have been reduced significantly (less number of replications), elaborately combined physician's schedule and patients' schedules would give a more comprehensive and better solution to ASFs.

CHAPTER 7

SUMMARY & FUTURE RESEARCH

Ever since the change of the healthcare reimbursement policy, medical practitioners are required to offer more effective and efficiency medical services to patients. To reduce the length of stay in hospitals, outpatient surgeries are increasing and ambulatory surgical facilities (ASFs) are widely open because it allows patients to finish their surgery within one day. The ASFs are either as the departments of one hospital or the stand alone facilities, depending one physician groups assigning patients to them, but both physicians and ASFs are paid by the private insurance companies or public insurance. However, all the medical bills relate to the number of patients brought by physicians and are not fully paid by insurance, so the operating cost for ASFs is becoming one big issue under such circumstances.

In addition, ASFs also wants to give the physicians good surgical environment and time schedules to make them attract more patients. Operating cost, physician's schedules, resources levels and quality of care are the most important factors for ASFs but none of the previous papers combined them all to study this ASF system. The multi aspect objective functions are set up includes the doctors' delay penalty cost, medical staffing resources delay penalty, patients delay penalty and staffing salary cost. Because of the complicity of the system, the absolute optimal results cannot be achieved but discrete-event simulation model has been built to evaluate different strategies for different levels of topics.

The first topic is about finding the optimal staffing strategy for ASFs which will results to a lower level of the combined total cost the authors mentioned above. After running the experiments, the conflicting matter between staffing salary cost and the other three penalties has been displayed and the overall lowest cost which will tradeoff all the factors are found, and it is not sharp convex but a U-convex, which means best staffing strategies can be achieved within a tight but flexible area.

The next topic is the study of physician groups' scheduling problem, five heuristic algorithms have been generated and tested on ten environmental problems, and these ten environmental problems are randomly set up with different parameters to stand for different scenarios of ASFs. Linear programming with balancing the resource usage objective is used to reduce the medical staffing's delay penalty. Among these five heuristic algorithms, heuristic #4 which is the operative and resource balancing algorithm gave the best results for nine problems. The lower bound for each problem is also simulated and all the best results have shown a small gap (20%) compare with the lower bound. All the comparisons are based on statistical analysis results, and scheduling solutions based heuristics have shown significantly better results than the dummy schedules. Though not with any absolute optimal strategy, by balancing the operative and resource usage gave some inspirations to further studies.

In the physicians' scheduling topic, the patient arrival ratios are provided as input to the simulation model which may arose higher variance in patients' arrivals. To have a more specific time for patients' arrival, patient individual scheduling has been studied by referring to flow-shop problem. Classical three machine minimizing makes span heuristics have been borrowed to apply in this three process problem.

Though not the exact same situations, the applications of these classical heuristics have a significant effect on the same ten problems. Seven heuristic algorithms are introduced and five heuristic algorithms are compared. There is no dominance among the five heuristics but the total time SPT rule and Palmers rule gave the better results for four problems. After separate cost analysis, proof is found to support that SPT rule has reduced patient waiting penalty effectively mean while palmer's rule has decreased the overflow penalty for staffing.

This ASF research has offered continuous solutions to ASFs' current problems via multiple objectives. By using discrete-event simulation model, the initial (staffing levels and physician, patient scheduling) decision making module has created, and the heuristic algorithms for physicians' scheduling problem and individual patient arrival problem have been proved effective to the current ten problems.

7.1. ASF Trends and Managerial Insights

In addition to the healthcare systems engineering research community, the primary group for whom this research is of significance is ASF managers and operators. The long term expectation is that ASFs will transition from a more manual expertise based decision making to model based data drive decisions making. There are several trends affecting the ASF business that will drive this trend, the authors list these first.

- Progressive (inflation adjusted) decrease in surgery compensations rates for both physicians and ASFs.
- Expansion of shorter surgery time procedures in the ASF portfolio.
- Transition of more procedures currently restricted to hospital ORs to ASFs. These procedures will more resource intensive relative to current portfolio.

- Progressive growth in the number of larger (6+ ORs) corporate owned ASFs catering to a much large set (8+) of physician groups.
- Large ASFs negotiating preferential pricing and exclusive service arrangements with insurance companies.
- Decline in small physician owned ASFs.
- Capacity growth providing physician with more choices in terms of where patients are directed.

The results of this research provide several ASF managerial insights which are highlighted below.

The practice of looking only at the direct staffing costs as the ASF operations planning objective is short sighted. The objective needed to be expanded to include quantitative assessments of the physician delay and patient delay costs. Use of the reliable costs coefficients presented here allows ASF management to build a competitive position in surgery practice industry. Traditional practice of simply basing decisions on physician and patient surveys is not sufficient.

Simulation models can provide accurate estimates of staffing overtime, physician delay and patient delay for a given staffing level and patient load. ASF management should and can use these models to make better decisions which optimize their overall operating performance. These models become even more critical as the variety of surgeries and the resource use complexity increases. The research has shown in many cases the optimal decision is quite distinct hence even small deviations can result in significant performance drops. These behaviors cannot be reliably estimated using just human experience.

The physician block assignment problem formalized here introduces a key decision making model for ASF managers. Certainly, the current practice of making assignments simply on relationships with physician groups should be replaced with analytical solutions methods developed here. This allows optimal combinatorial data driven solutions to be developed.

The patient scheduling problem formalized and introduced here should replace the current practice of scheduling batch patient arrivals on a fixed interval. These models base the arrival decisions on the current physician schedule and the mix of surgeries scheduled for a given day. The model developed here will allow ASFs to better utilize current communication technology to provide patients with day before arrival schedule that minimize their patient delay. As this research as shown due to the cost ratios between physician and patient delay coefficients, patient delays tend to be a secondary objective. The use of the patient scheduling model will help mitigate this situation.

7.2. Future Research Plan

However, because of some limitations, there is still some work to be done in the future. Even though from the staffing level optimization, a prediction for levels of staffing is provided, an imbalanced medical practitioners from geographically and occupations as well would result to human resource problems. The limitation of implementing, even the smart strategies are offered by the research, the scheduling priorities from the physicians' side are more difficult than theoretic planning.

There is some percentage of no show up or cancelations in the ASFs, which will cause resource waste and different scenarios are under discussed for that case. The above

limitations are exist in ASFs but are uncontrollable from simulations, or mathematic programming. The ASF Simulation model developed here along with the accompanying problems solved provides a rich platform for future research. Specifically, complex ASF features and parameter changes can be studied.

1. Other non-linear form and physician/patient priority of the objective function.
2. Integrated heuristics for solving the Physicians Block Assignment plus Patient Arrival Scheduling problem.
3. Better heuristic for Physician Block Assignment and Patient Arrival Scheduling problem.
4. Investigate sensitivity to different surgery time distributions.
5. Investigate the effect of ϕ_D and ϕ_P interaction on performance.

REFERENCES

1. Alexopoulos, C., Goldsman, D., Fontanesi, J., Kopald, D., & Wilson, J. R. (2008) Modeling patient arrivals in community clinics. Omega 36:33-43.

2. Alexopoulos, C., D. Goldsman, et al. A Discrete-Event Simulation Application for Clinics Serving The Poor."association, A. s. c. ASFs - A Positive Trend in Health Care.

3. Belien J and Demeulemeester E (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy. Eur J Opl Res 176: 1185–1204.

4. Cardoen, B., E. Demeulemeester, et al. (2010). "Operating room planning and scheduling: A literature review." European Journal of Operational Research 201(3): 921-932.

5. Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: A review of literature. Production and Operations Management, 12(4), 519–549.

6. Cayirli, T, E. Veral, and H. Rosen. (2006). Designing appointment scheduling systems for ambulatory care services. Health Care Management Science 9, 47–58.

7. Carter, D. L. a. D. M. (2010). Scheduling According to Physician Average Procedure Times in Endoscopy Suites. POMS 21st Annual Conference.

8. Denton, B., J. Viapiano, et al. (2006). "Optimization of surgery sequencing and scheduling decisions under uncertainty." Health Care Management Science 10(1): 13-24.

9. Denton, B., D. Gupta. 2003. A sequential bounding approach for optimal appointment scheduling. IIE Trans. 35(11) 1003–1016.

10. Denton, B. T., A. S. Rahman, et al. "Simulation Of A Multiple Operating Room Surgical Suite.."

11. Dexter F, Epstein RH, De Matta R, Marcon E (2005) Strategies to reduce delays in admission into a postanesthesia care unit from operating rooms. (Journal of PeriAnesthesia Nursing 20(2)) 92–102.

12. Dexter F, Ledolter J. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historical data. Anesthesiology. 2005;103;1259-1267

13. Cooper, K., Brailsford, S. C., & Davies, R. (2007) Choice of modelling technique for evaluating health care interventions. Journal of the Operational Research Society 58: 168-176.

14. Kumar, A., & Shim, S. J. (2005) Using computer simulation for surgical care process reengineering in hospitals. INFOR 43:303-319.

15. Durant, G. D. (1993). "Expanding the scope of ambulatory surgery in the USA." Ambulatory Surgery 1(4): 173-178.

16. Durant, G. D. and C. J. Battaglia (1993). "The growth of ambulatory surgery centres in the United States." Ambulatory Surgery 1(2): 83-88.

17. Erdogan, S. A. and B. Denton (2011). "Dynamic Appointment Scheduling of a Stochastic Server with Uncertain Demand." INFORMS Journal on Computing.

18. Ferreira, R. B., F. C. Coelli, et al. (2008). "Optimizing patient flow in a large hospital surgical centre by means of discrete-event computer simulation models." Journal of Evaluation in Clinical Practice 14(6): 1031-1037.

19. Franklin Dexter, M., PhD*, and Rodney D. Traub, PhD† (2002). "<How to Schedule Elective Surgical Cases into Specific.pdf>."

20. Franklin Dexter, M., PhD, Alex Macario, MD, MBA, Rodney D. Traub, PhD, and P. Margaret Hopwood, and David A. Lubarsky, MD "<An Operating Room Scheduling Strategy to Maximize the.pdf>."

21. Gul, S., B. Denton, et al. "Bi-Criteria Scheduling Of Surgical Services For An Outpatient Procedure Center"

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

23. Hair, B., P. Hussey, et al. (2012). "A comparison of ambulatory perioperative times in hospitals and freestanding centers." The American Journal of Surgery 204(1): 23-27.

24. Hollingsworth, J. M., S. L. Krein, et al. (2010). "Opening Ambulatory Surgery Centers and Stone Surgery Rates in Health Care Markets." The Journal of Urology 184(3): 967-971.

25. Hsu, V. N., R. de Matta, et al. (2003). "Scheduling patients in an ambulatory surgical center." Naval Research Logistics 50(3): 218-238.

26. Joshi, G. P. and R. S. Twersky (2000). "Fast tracking in ambulatory surgery." Ambulatory Surgery 8(4): 185-190.

27. Klassen, K. J., T. R. Rohleder. 2004. Outpatient appointment scheduling with urgent clients in a dynamic, multi-period environment. Internat. J. Service Indust. Management 15(2) 167–186.

28. Linda R. LaGanga, L.R and Lawrence, S.R., 2007, Clinic Overbooking to Improve Patient Access and Increase Provider Productivity, Decision Sciences, Volume 38 Number 2, May 2007

29. Lamiri, M., F. Grimaud, et al. (2009). "Optimization methods for a stochastic surgery planning problem." International Journal of Production Economics 120(2): 400-410.

30. Liu, N., Serhan Ziya, S. and Kulkarni, V.G,, 2010, Dynamic Scheduling of Outpatient Appointments Under Patient No-Shows and Cancellations, Manufacturing & Service Operations Management 12(2), pp. 347–364

31. Liu, L. and X. Liu. (1998). Block appointment systems for outpatient clinics with multiple doctors. Journal of the Operational Research Society 49, 1254-1259.

32. Marcon, E. and F. Dexter (2006). "Impact of surgical sequencing on post anesthesia care unit staffing." Health Care Management Science 9(1): 87-98.

33. Marcon E, Kharraja S, Smolski NN, Luquet B, Viale JP (2003) Determining the number of beds in postanesthesia care unit: a computer simulation flow approach, Journal of the International Anesthesia research society, (Anesthesia & Analgesia 96) 1415–1423

34. Macario A. Truth in scheduling? Anesth Analg. 2009;108:681-685, Medscape Anesthesiology,

35. Macario A. Is It Possible to Predict How Long a Surgery Will Last?, July 2010, http://www.medscape.com/viewarticle/724756

36. Noon, C. E., Hankins, C. T., & Cote, M. J. (2003) Understanding the impact of variation in the delivery of healthcare services. J Healthc Manag 48: 82-98.

37. Ogulata SN and Erol R (2003). A hierarchical multiple criteria mathematical programming approach for scheduling general surgery operations in large hospitals. J Med Syst 27: 259–270.

38. Reis, E. D., F. Mosimann, et al. (1999). "Implementation of ambulatory surgery in a university hospital: an audit comprising 873 general surgery cases." Ambulatory Surgery 7(2): 107-110.

39. Roberts, L. (1994). "Accreditation of ambulatory surgery centres utilizing universally acceptable clinical indicators — is it achievable?" Ambulatory Surgery 2(4): 223-226.

40. Rohleder, T. R., & Klassen, K. J. (2002). Rolling horizon appointment scheduling: A simulation study. Health Care Management Science, 5(3), 201–209.

41. Rohleder, T. R., Bischak, D. P., & Baskin, L. B. (2007) Modeling patient service centers with simulation and system dynamics. Health Care Manag Sci 10:1-12.

42. Spangler WE, Strum DP, Vargas LG and May JH (2004). Estimating procedure times for surgeries by determining location parameters for the lognormal model. Health Care Mngt Sci 7: 97–104.

43. Strum DP, May JH and Vargas LG (2000). Modelling the uncertainty of surgical procedure times: comparison of log-normal and normal models. Anesthesiology 92: 1160–1167.

44. Strum, David P., Jerrold H. May, and Luis G. Vargas. "Modeling the Uncertainty of Surgical Procedure Times: Comparison of Log-Normal and Normal Models." Anesthesiology 92, no. 4 (2000): 1160-1167.

45. Strum, D., J. May, A. Sampson, L. Vargas, and W. Spangler (2003). Estimating times of surgeries with two component procedures. Anesthesiology 98, 232{240.

46. Thor, J., J. Lundberg, et al. (2007). "Application of statistical process control in healthcare improvement: systematic review." Qual Saf Health Care 16(5): 387-399.

47. Vasilakis C, Sobolev BG, Kuramoto L and Levy AR (2007). A Simulation study of scheduling clinical appointments in surgical care. J Opl Res Soc 58: 202–211.

48. Weng, M. L. and A. A. Houshmand "Healthcare Simulation A Case Study At a Local Clinic."Wikipedia.from http://en.wikipedia.org/wiki/Arena_(software).

49. Walley, P., Silvester, K., & Mountford, S. (2006) Health-care process improvement decisions: a systems perspective. International Journal of Health Care Quality Assurance, 19: 93-104.

50. Yeung, Y. P., F. L. Cheung, et al. (2002). "Survey on postoperative pain control in ambulatory surgery in Hong Kong Chinese." Ambulatory Surgery 10(1): 21-24.

51. Young, T. (2005) An agenda for healthcare and information simulation. Health Care Manag Sci 8:189-196.

52. Zhang, B., P. Murali, et al. (2008). "A mixed integer programming approach for allocating operating room capacity." Journal of the Operational Research Society 60(5): 663-673.

53. Zhu, Z. C. H., B.H. &Teow, K. L. (2009). "Estimating ICU bed capacity using discrete event simulation." International Journal of Simulation Modelling 3: 10.

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