OPTIMIZING SURGICAL SCHEDULING

Through Integer Programming and Robust Optimization

Supervisor: Prof. Jonathan Patrick

Co-Supervisor: Prof. Jonathan Li

Telfer School of Management

University of Ottawa

By :

Shirin Geranmayeh

Abstract

This thesis proposes and verifies a number of optimization models for re-designing a master surgery schedule with minimized peak inpatient load at the ward. All models include limitations on Operating Rooms and surgeons availability. Surgeons` preference is included with regards to a consistent weekly schedule over a cycle. The uncertain in patients` length of stay was incorporated using discrete probability distributions unique to each surgeon. Furthermore, robust optimization was utilized to protect against the uncertainty in the number of inpatients a surgeon may send to the ward per block. Different scenarios were developed that explore the impact of varying the availability of operating rooms on each day of the week. The models were solved using Cplex and were verified by an Arena simulation model.

Table of Contents

List of Figures

List of Tables

Introduction

One of the common practices in hospitals is to organize their operation theatres' schedules in order to maximize throughput. Operating Rooms (ORs) are one of the most expensive resources and their efficient utilization is of high importance for hospital managers. On the other hand, maintaining the occupancy rate in wards at a manageable level (thus ensuring the responsiveness of the hospital to natural uncertainties with regards to arrival rates and lengths of stay) is also of high concern to managers. Clearly, the operating rooms' utilization has a direct effect on bed utilization in the wards, but conversely, high bed utilization may have a negative effect on ORs` utilization as high occupancy rates in the wards may lead to delays post-surgery or even surgery cancellations. Both OR utilization and bed occupancy have been investigated in several research papers. However, there are few studies addressing the combination of the two. In our study, we wish to investigate the effect of scheduling operating rooms on utilization and bed occupancy at the ward level. The locus for this study is The Queensway Carleton Hospital (QCH). QCH is a community hospital in west Ottawa which offers a diversity of medical and surgical services. Equipped with 264 beds and over 1800 healthcare professionals, QCH is the secondary referral center for the Ottawa Valley. The surgical unit of the hospital includes 56 beds but surgical demand for beds often exceeds this threshold. The overflow of surgical patients is managed by using internal medicine beds. Serving the overflow of surgical patients by occupying medical beds causes delays for patients seeking access to the medical unit. Delaying service for a critical emergency patient due to insufficient bed capacity risks the patients` safety.

Due to the uncertain and volatile nature of emergency arrivals to the hospital, the scheduling of elective surgeries is considered as a means of managing the patient load at the hospital. Therefore, QCH is interested in studying the patient flow into and out of the operating rooms in order to optimally manage bed occupancy in the surgical ward.

More than 45 surgeons provide surgical services in 6 different specialties at QCH. These 6 specialties are: general surgery, orthopedic surgery, plastic surgery, gynecology and obstetrics, otolaryngology (Ear, Nose, and Throat- ENT) and urology. Surgeries are run in 8 out of 11 available operating rooms. However, even with 3 dormant operating rooms, the QCH surgical unit is the third largest surgery unit in Ottawa serving both elective and emergency cases. In total, over 10,000 day surgeries and inpatient procedures take place in QCH in a year.

Patients- whether emergency, elective or medical- are categorized according to the type and amount of resources they need in the hospital. This classification is called the Case Mix Group (CMG). The standard for this classification is developed based on data of acute care inpatient characteristics and cost records. QCH covers services for over 500 CMGs. In addition to CMG, QCH utilizes a secondary patient classification system with a five point scale depending on the severity of their co-morbidities. These two classifications help hospital management better estimate cost indicators such as patients` Length of Stay (LOS).

The master surgery schedule includes a 4 week cycle of surgical blocks. Each block is allocated to a specific specialty and to a specific surgeon of the assigned specialty. Each surgeon schedules surgeries for his own assigned blocks with a restriction on the sum of the average length of all procedures being less than the length of the block.

Patients are grouped into two main categories: 1) same day stays are those who are discharged from the hospital on the day of surgery, and therefore require no surgical bed and 2) inpatients that need a longer stay at the hospital before discharge. Clearly not all surgical blocks generate inpatients for the hospital and the actual number of inpatients generated by each surgeon may vary from one block to the next.

Currently, demand for surgical beds in QCH outpaces availability. This is resolved through "borrowed" beds meaning that surgical patients are cared for in internal medicine beds. Conversely, when demand for surgical beds is below capacity (often on the weekends), internal medicine patients from the emergency department are placed in surgical beds. QCH management is interested in determining if there is a way to reduce these off-service placements through improved elective patient scheduling. The actual scheduling of patients into the blocks is under the purview of the surgeon and thus not readily influenced by hospital management. On the other hand, the allocation of blocks to surgeons is the responsibility of hospital management. Thus QCH management is interested in re-designing the surgical master schedule at the level of the surgeon and in such a way as to minimize the peak load bed requirement and thus reduce the overflow of patients into medical beds.

Research questions

Blake et al (Blake JT, Dexter F, & Donald J , 2001) was one of the first to determine that the variation in ward occupancy could be managed and smoothed by improved scheduling of the elective stream of arrivals to the ward. Our data analysis suggests that elective occupancy represents a cyclic pattern, in which, the ward faces the highest occupancy at mid-week and the lowest occupancy on the weekends. Thus the initial hypothesis was to test whether that variation could be mitigated (and by how much) through a manipulation of the master schedule. This project allowed us to address the following questions:

1. What master surgery schedule can provide the surgical ward with optimized bed occupancy while maintaining a similar throughput to current practice at QCH? What would be the effect of opening ORs on the weekends?

2. How does the incorporation of the variability in length of stay according to each surgeon affect the optimal master schedule?

3. How does providing surgeons with a consistent schedule from week to week affect the optimal schedule?

4. How can we incorporate the variability in the number of inpatient referrals each day by each of the surgeons into the model? How might this affect the optimal block schedule?

5. How can we safeguard the ward against periods of congestion?

The above questions are addressed using Mixed Integer Programming and Robust Optimization. Research questions 1 to 4 are addressed in Part A of the Methodology section where we develop an Integer program and extend it to meet the preferences of surgeons. In this part, we analyze the results of different scenarios to evaluate the effect of different parameters on a surgical unit`s performance measures. Moreover, further insight on the output of our model and its consequences on bed utilization are provided in Part B by developing a simulation model that incorporates emergency surgical patients and medical patients as well as elective surgical patients. Part C covers research questions 5 and 6 by applying Robust Optimization.

Literature Review

Capacity management at hospitals has often been studied in a compartmental fashion where the capacity requirement of a single unit such as the Emergency Department (ED), ward, lab, OR or diagnostic facility is ″optimized″ in isolation. However, due to the interconnected nature of hospital activities and procedures, results of studying one isolated subsystem may result in undesired consequences for other connected but ignored subsystems. For example, at the hospital level, Lane et al developed a system dynamics model of the emergency department to examine the sensitivity of waiting times to hospital bed numbers (Lane, Monefeldt, & Husemann, 2003), (David C. Lane, Monefeldt, & Rosenhead, 2000). At the time, the British government believed that improved efficiency could be achieved by decreasing the number of beds. The rationale was that reduced beds with the same level of demand will increase the use of the remaining beds thereby increasing efficiency and hospital services will compensate for the reduced number of beds by working harder. However, the model demonstrated that this was not the case. Instead there would be an increase in the cancellation of elective surgeries. The system was able to handle an increase of up to 4% in demand. Beyond 4%, the waiting times rose steeply. In the event of a crisis (e.g. a 13% increase in demand for 24 hours), the system took an additional five days on top of the crisis day to recover. This study demonstrates how focusing on a single issue can simply transfer the problem elsewhere in the system.

On the other hand, a systemic and holistic approach to the analysis of a health system is often sufficiently complex as to be intractable. Thus, we aimed to build a model that is complex enough to be useful but simple enough to remain tractable. Since surgical

departments usually contribute greatly to the hospital`s total expenditure and level of service (McCrory, LaGrange, & Hallbeck, 2014), in this study, we include two sections of QCH - ORs and the surgical ward- and investigate the consequences of actions in one on the other. However, we keep in mind that there may be other consequences on other sections of the hospital (such as the medical ward, ED, diagnostic facilities, clinics) that are not incorporated into the model but that we incorporate through simulation.

In this section, first we review research on the scheduling of operating rooms and focus on those who approached OR scheduling according to the postoperative resources availability. Our initial interest is the application of Mixed Integer Programming in scheduling. We investigated each paper to determine how the authors defined the objective function and the constraints and how they approach solving and validating their model.

In a later step, we investigated the possibility of applying Robust Optimization (RO) with the goal of incorporating another level of stochasticity in our model. RO seeks a certain level of robustness and protection against uncertainty of parameters or cost coefficients. We would explain more about RO in the following sections. In the literature review we focus on the application of Robust Optimization to healthcare applications. With RO research, we are not only interested in the definition of the objective function and constraints but also with the uncertainty sets implemented.

Gupta (Gupta, 2007) investigated three main problems in hospitals with regards to operating theatres: capacity allocation for elective surgeries, elective surgery booking control and sequencing surgeries in each surgical block. The first two classes of problems are similar to what we include in our model. In the first phase, Gupta`s linear optimization model deals with demand per specialty and available capacity allocated to

each specialty. The objective function is set to maximize the total contribution (revenue) to the hospital by optimizing the capacity allocation to different specialties subject to the constraint that each specialty maintains a minimum in order to maintain clinical and economic viability. The second phase describes a model for surgery booking control. The only constraint is to consider downstream critical resources such as bed and ICU capacity since they may cause blockages in the OR, overtime or long waiting times. He introduces urgency, patient characteristics and revenue as criteria to classify patients. He models this problem on a daily basis using a rolling horizon. His model can be viewed as an aggregate planning model for downstream resources.

Pham and Klinkert (Pham & Klinkert , 2008) developed a Mixed Integer Programming (MIP) model that is adapted from a job shop scheduling problem to schedule individual elective surgeries. The model helps determine the best time to perform each surgery. They equated each surgical case to a component in industry that can be processed by different types (sets) of resources with different availability intervals through each step. They defined three steps corresponding to the preoperative, perioperative and postoperative time periods based on which they adapted the Multi-Mode Blocking Job Shop (MMBJS) approach to minimize the makespan of an elective surgery patient and at the same time to force the operation to be scheduled as soon as possible. They solved the model using Cplex solver.

Min and Yih (Min, D. & Yih, Y., 2010) address OR scheduling including variability in the duration of different procedures and the availability of downstream resources over periods. The problem objective is to minimize both patient cost and overtime cost. Patient cost is associated with delays in treatment and waiting time such as disease progression, pain or dysfunction and disability. They use a MIP model at the block scheduling level.

They adopted the sample average approximation technique in order to deal with the stochastic nature of their optimization model. They verified their results by a simulation model.

Shylo et al (Shylo O.V, Prokopyev O. A, & Schaefer A. J, 2013) use MIP to maximize the utilization of the operating room. They address the OR scheduling problem at the individual surgery level and not at the OR master schedule. They develop a batch scheduling framework to book a set of surgeries into an ordered set of available blocks. Block booking is mainly concerned with the balance between block utilization and block overtime. They approximate the sum of procedure durations to a normal distribution and provide near-optimal solutions for stochastic scheduling and show that batch scheduling has better performance rather than open booking (sequential booking). Open booking books the first surgery case arrival to the first available and appropriate (in case of specialty, time available, etc.) slot.

Olivares et al (Olivares M, , Terwiesch C, & Cassorla L, 2008) developed an economic model to balance reserving too much versus too little OR capacity for a specific specialty (cardiac surgery). Reserving too much capacity results in idle time costs and reserving too little capacity results in overtime costs. According to the context of their research and the available data, they conclude that the hospital under investigation was placing too much emphasis on idle time as the implied cost of idle time was 60% higher than the overtime cost. They clarify that this cost allocation only reflects the hospital's current strategy in aligning conflicting objectives. They characterize their model as a newsvendor model and apply cost ratio analysis (underage vs. overage cost) to provide a suitable suggestion for the hospital.

Astaraky and Patrick (Astaraky & Patrick, 2015) address multiple objectives by minimizing a weighted sum of patient waiting time, overtime in the OR and congestion in the ward. They model the problem using a Markov Decision Process model and apply Approximate Dynamic Programming to solve the model. They use the K-means clustering technique to classify patients from over 900 surgical procedures across 9 specialties into a manageable number of subgroups based on procedure length in the OR and post-operative LOS in the ward. The authors used the newly defined subgroups to book patients within a pre-defined master block schedule. Thus, their model is concerned the booking of patients into blocks rather than the scheduling of the blocks themselves. They verified the performance of their model by simulation.

Blake et al (Blake JT, Dexter F, & Donald J , 2001) used Integer Programming to allocate different specialties with different target times to surgical blocks to minimize the shortfall across the specialties. They used the estimated demand for each specialty and developed a methodology for allocating a block of a certain OR in a certain day of the week to each specialty in order to best meet the demand across specialties. They came to the conclusion that although the master schedule is a cyclic schedule, it must be flexible among cycles to better meet the demand.

Van Oostrum et al (Van Oostrum J.M.x, Van Houdenhoven M., & Hurink J.L., 2008) develop a mathematical program with probabilistic constraints in order to best utilize OR capacity without increasing overtime or cancellations while levelling bed occupancy in the ward and intensive care units. They dealt with the uncertainty of procedures using planned slack to avoid the probability of overtime. However, to minimize the required slack, they used the portfolio effect principle which measures the tendency of risk for a portfolio of stochastic variables and accounts for variation cancellation among variables

in a particular portfolio. They considered different surgery procedures` duration as a portfolio of stochastic variables and tried to minimize the risk of overtime and minimize the required slack. They developed a heuristic approach to solve their min-max objective function which is based on an integer linear program with probabilistic constraints. They tried to minimize the maximum demand for hospital beds during a scheduling cycle. Authors proposed a column generation method to solve this NP-hard problem.

Carnes and Price (Carnes, T & Price D. , 2011) investigate after-surgery patient flow at a hospital through the Post Anaesthesia Care Unit (PACU) and surgical inpatient beds. They applied integer programming in order to rearrange the OR master schedule at the surgeon level to reduce peak occupancy downstream. Authors consider OR specifications in assigning specialties to each. Moreover, they consider constraints that support the need for more than 1 OR to be working together (linked ORs) for specific types of surgeries such as transplant. They considered both midday and midnight census to evaluate the permuted master schedule against the current one. By permutation of blocks they were able to reduce the peak load of inpatient beds by 28% for the midweek which represented the highest occupancy in the weekly cycles throughout the year.

Santibanez et al (Santibanez, P., M. Begen, & D. Atkins , 2007) developed a MIP model to schedule surgical blocks for multiple hospitals with different available specialties at each hospital. They grouped procedures together in each specialty and within each hospital, and computed parameters such as average duration of surgical procedures per group and average required resources after surgery. The model was then used to determine how many procedures can be performed in each surgical block in any of the hospitals. They address multiple objective functions including minimizing the sum of the maximum usage of post-surgical resources and beds- in line with our objective for QCH.

In a study following Santibanez et al, Chow et al (Chow, Puterman, Salehirad, Huang, & Atkins, 2011) combine a Monte Carlo simulation model and mixed integer programming. They used simulation to predict the bed demand for any given surgical block schedule, and then used MIP to generate improved surgical block schedules. They sought to minimize the peak demand across the wards rather than the total or average demand. The logic of grouping patients is different from Santibanez et al.`s study. Santibanez et al.`s groups are based on medical and surgical procedures while Chow defines the groups according to patients length of stay. To smooth bed demand, Chow et al. (2011) recommend that procedures with long Estimated Length Of Stay (ELOS) should be scheduled on Mondays to (smooth) optimize bed demand during the week, or on Fridays to increase bed demand during the weekend.

Beliën and Demeulemeester (Beliën & Demeulemeester , 2007) introduce three stages for operating room scheduling: choosing the specialties that will be covered by a hospital (referred to as the case mix), developing a master surgery schedule which assigns blocks of ORs to specialties and finally the scheduling of individual procedures into blocks on a daily basis. They develop a master schedule by mixed integer programing solved by heuristics and meta-heuristic methods aimed at minimizing the expected bed shortage by leveling bed occupancy. Their problem is proved to be NP-hard so they use a repetitive Mixed Integer Programming heuristic by adding an extra constraint to the model after successive iterations. In another attempt, they try a heuristic approach for a quadratic objective function. They include two types of constraints: demand constraints and supply constraints in which bed capacity is a constraint on the supply side. The number of inpatients redirected to the ward and the length of stay in the ward are both considered as stochastic variables following multinomial distributions.

In 2009, Beliën and Demeulemeester (Beliën. J & Demeulemeester, E. , 2009) embedded the developed their earlier model in a decision support system in order to design a cyclic master schedule. The initial objective of the decision support system is to level bed occupancy as much as possible. Moreover, other objectives that the decision support system fulfills are 1) to assign the same OR to the same specialty as much as possible, and 2) to make the cyclic master schedule as monotone as possible with minimal changes between weeks. To measure the performance of this multi-objective model, they define a weighted penalty function which measures the quality of each developed scenario by the system and represents the cost of sacrificing the performance of one of the objectives in favor of improving another. The developed mixed integer linear and quadratic objective functions are solved by Cplex and by heuristic approaches.

In our study, we initially focus on the Belien and Demeulemeester 2007 and 2009 papers. Both studies address similar issues that QCH is already facing such as OR availability, bed capacity, surgeons' privileges and stochastic LOS. The main method used is MIP. In the second stage we incorporate more stochasticity into our model with regards to the number of inpatient referrals per block per surgeon. We consider Robust Optimization (RO) as a method to include this variability to protect the decision maker against parameter ambiguity and stochastic uncertainty. Robust optimization is a relatively new method in optimization which helps guard against the worst case scenario; where the solution is evaluated using the realization of the uncertainty that is most unfavorable. RO is used in different fields such as inventory management, dynamic pricing, portfolio management, etc. There are also some limited applications in the healthcare literature where RO has been used primarily to aid clinical decisions rather than managerial ones.

For instance, Bortfeld et al (Bortfeld, T., Chan, T.C.Y., Trofimov, A., & Tsitsikl, J.N, 2008) developed a robust optimization model to optimize intensity-modulated radiation therapy. They built a model of motion uncertainty to describe breathing motion using a probability density function. Breathing motions during radiation disrupt dosage delivery to the targeted organ. Their robust model provides a solution with lower levels of under dosage and higher protection for surrounding organs.

Chat and Mišić (Timothy C.Y. Chan & Velibor V. Mišić , 2013) extended the previous work and developed a Dynamic Robust Optimization model and solved a sequence of RO models in which the uncertainty set is updated according to the information gathered in previous treatment sessions for each of the iterations. This method generally shows better solutions than static RO models. It also suggests that the initial uncertainty set is less important than in the static RO for the final solution.

Holte and Mannino (Holte M. & Mannino C. , 2012) developed a RO model to develop a master surgery schedule in order to minimize the queue cost for surgery with regards to uncertainty in weekly demand for each specialty. Their objective was to design a safe cyclic schedule which performs optimally for the worst case scenario meaning that the queue size is minimized for scenarios where the demand is highest. They applied their model to design the schedule for a hospital in Oslo.

Addis et al (Addis, Carello, & Tànfani, 2014) applied RO to the Advance Scheduling Problem by which they determined which cases should be allocated to each surgical block in a given planning horizon with a defined set of ORs and a waiting list for surgeries. The source of uncertainty in their model is procedures` duration and the objective is to minimize a measure of waiting time for elective cases which concerns urgency and tardiness for the patients. The performance of the solution is evaluated for

OR utilization by random scenarios for lognormal stochastic surgery procedures and random cancellations.

Denton et al (Denton, Miller, Balasubramanian, & Huschka, 2010) developed a combinatorial optimization model in order to schedule individual procedures into ORs on a given day of surgery. The nominal model is a two-stage stochastic linear program with binary decision variables which minimizes the overtime cost. In the robust extension of their model, they considered the duration of surgical procedures as the source of uncertainty and tried to minimize the maximum cost associated with this uncertainty. According to their numerical results, they concluded that the robust model performs nearly as good as the heuristic methods they applied for the stochastic model.

Rachuba and Werners (Rachuba & Werners, 2014) applied a multi-objective and multi-criteria mixed integer optimization modeling approach for block scheduling. They considered the interests of three categories of stakeholders in their modeling approach: patients, staff and management. Respectively, their objectives address waiting time, over time and number of treated patients as corresponding to each of the target stakeholders. They applied a robust approach in order to include uncertainty of elective procedure times or arrival and duration of emergency cases.

In our research, we use RO to incorporate the uncertainty regarding the number of inpatients resulting from each block in order to minimize the effect on maximum peak load over days. Rather than assuming a fixed number of inpatients for each surgeon, we follow the approach of RO and treat the number as a variable that can take any value within a pre-described set. The objective is to determine which value taken in the set will lead to the worst-possible maximum peak load over the cycle. According to RO, we can formulate this as a maximization problem over the number of inpatient referrals per block

per surgeon. This additional layer of maximization problem can be solved using duality theory that converts the maximization problem into a minimization problem which can be integrated into the initial MIP (as will be shown later). The solution generated from the RO model provides a schedule that performs well even during periods of congestions when the number of inpatient referrals is higher than the average. Furthermore, our model determines which surgeons cause the greatest disruption to ward occupancy when they exceed their average patient referrals in a block. To the best of our knowledge, our work should be the first in the literature that applies RO in surgical scheduling with the aim of managing post-surgery resource utilization.

Data Analysis

We had access to *summarized* data for more than 6500 surgical patients from QCH for the fiscal year 2012 to 2013 and more than 4300 surgery patients for the fiscal year 2013 to 2014. These data sets cover all surgical patients receiving surgery in the specified time frames as well as the specialty of services required, their pathway through the hospital and among the different units, their length of stay at each unit, attending surgeons, their medical classifications, type of required/provided medical services, etc. We also had access to surgeons' availability and the current master schedule.

First of all, we were interested in the census of elective patients in the ward on each day in order to understand the as-is situation. The current elective bed occupancy per day for each cycle according to the data of year 2013 is shown in Figure 1. As mentioned before, each scheduling cycle is 4 weeks in length.

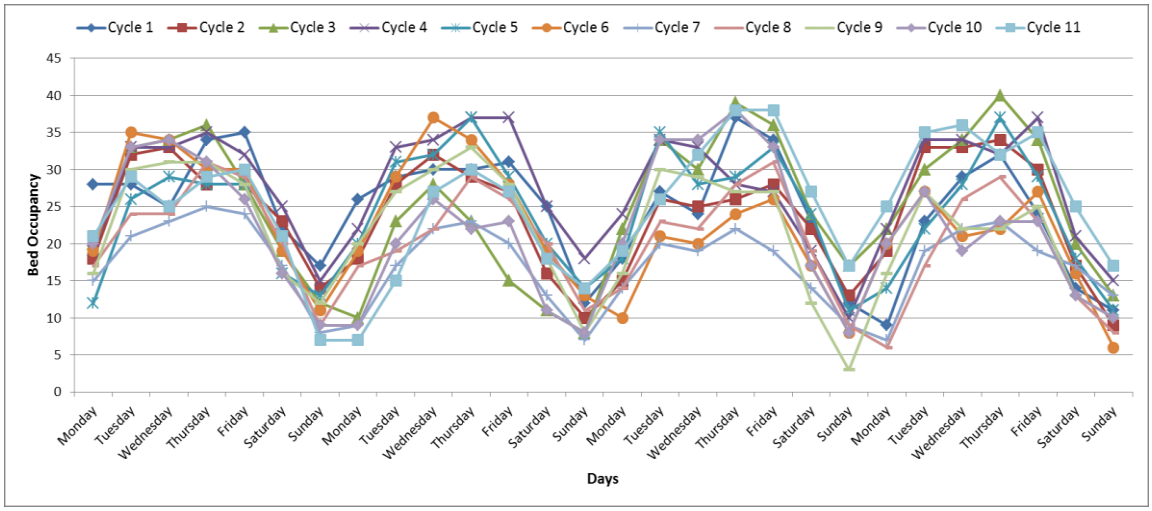

Figure 1: Bed census per cycle for elective cases- 02-01-2013 to 27-12-2013

Figure 2 represents the average occupancy by elective patients at ward over the cycles throughout the year.

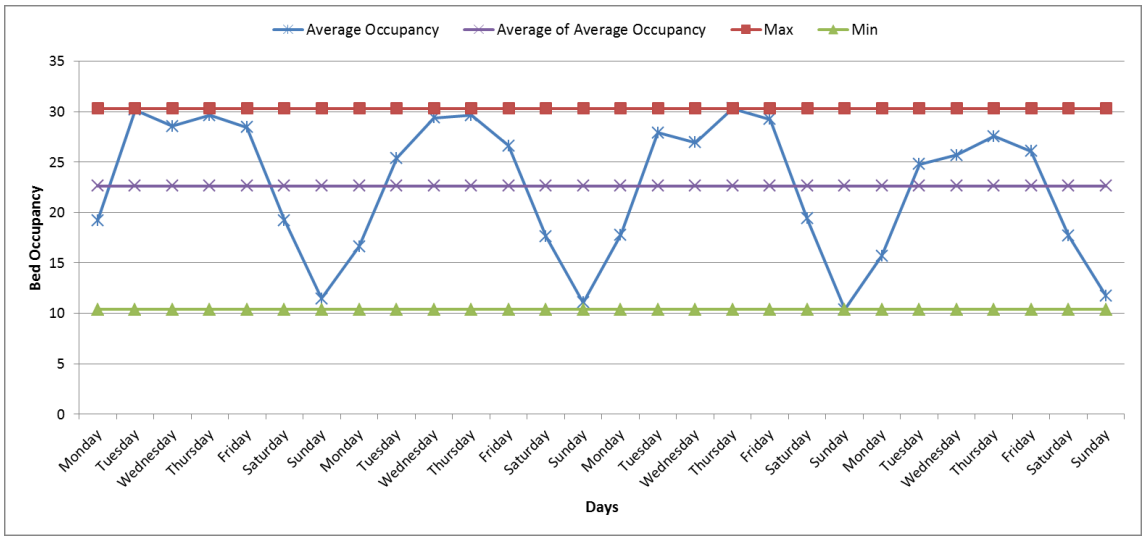

Figure 2: Average cyclic bed census for elective cases- 02-01-2013 to 27-12-2013

Figures 1 and 2 demonstrate the cyclic manner of the fluctuation in bed occupancy due to elective cases in the ward. The peak occupancy occurs mid-week while the ward is largely underutilized over the weekend. It is worth noting that the surgical ward is not only utilized by elective cases but also by emergency cases which is why the census of only elective patients does not demonstrate the overflow problem. Surprisingly, as demonstrated in Figure 3, the emergency census appears to be more stable than the elective one.

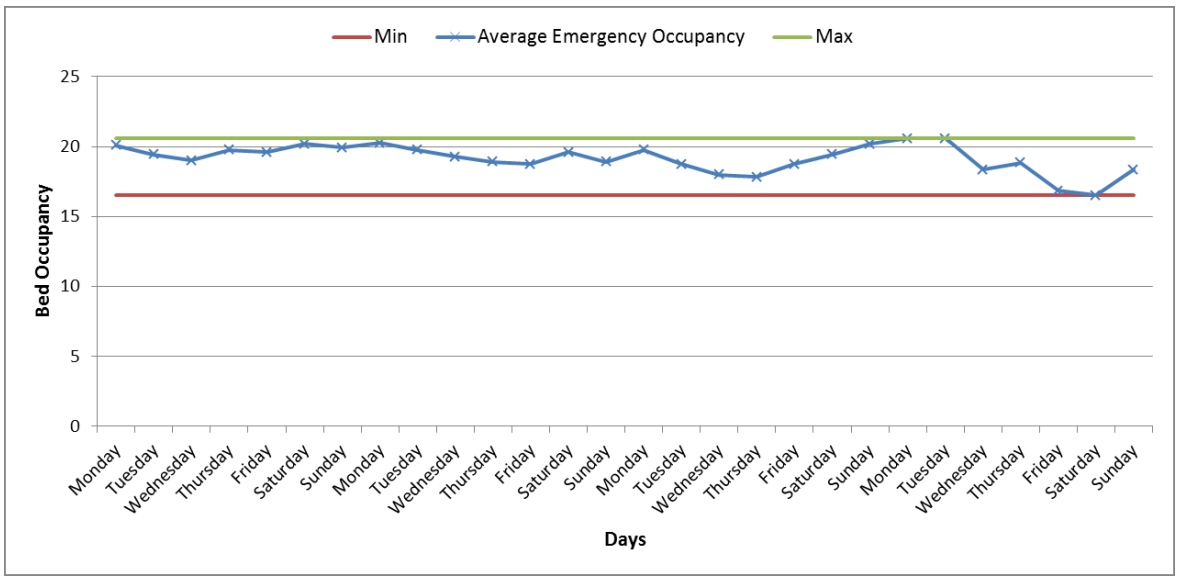

Figure 3: Average cyclic bed census for emergency cases- 02-01-2013 to 27-12-2013

Figure 3 shows the **average** emergency occupancy at ward over a cycle. Average minimum occupancy is 16.5 beds and average maximum emergency occupancy is 20.58 and the standard deviation is one.

We focused our analysis on surgeons who refer a significant number of inpatients to the hospital. Specifically, surgeons who operated and referred more than 30 inpatients to the ward in 2013 were included in the model. Other surgeons are aggregated as one auxiliary surgeon per specialty. Other surgeons were grouped together by specialty. Therefore, in our analysis we considered 23 individual surgeons plus 4 auxiliary (aggregated) surgeons, one each for general surgery, plastic surgery, gynecology and ENT. It should be noted that many of the surgeons under the umbrellas of auxiliary surgeons may deal more with same-day admission patients. Therefore, they are not included in our analysis since our focus is on bed occupancy in the surgical ward.

We also used the data to obtain LOS distributions per surgeon, procedure and specialty, as well as the patient referral to ward distributions for each surgeon. Each surgeon operates and refers a different number of inpatients to the ward per block. Figure 4 compares 23 surgeons based on the ratio of different number of inpatients generated per block.

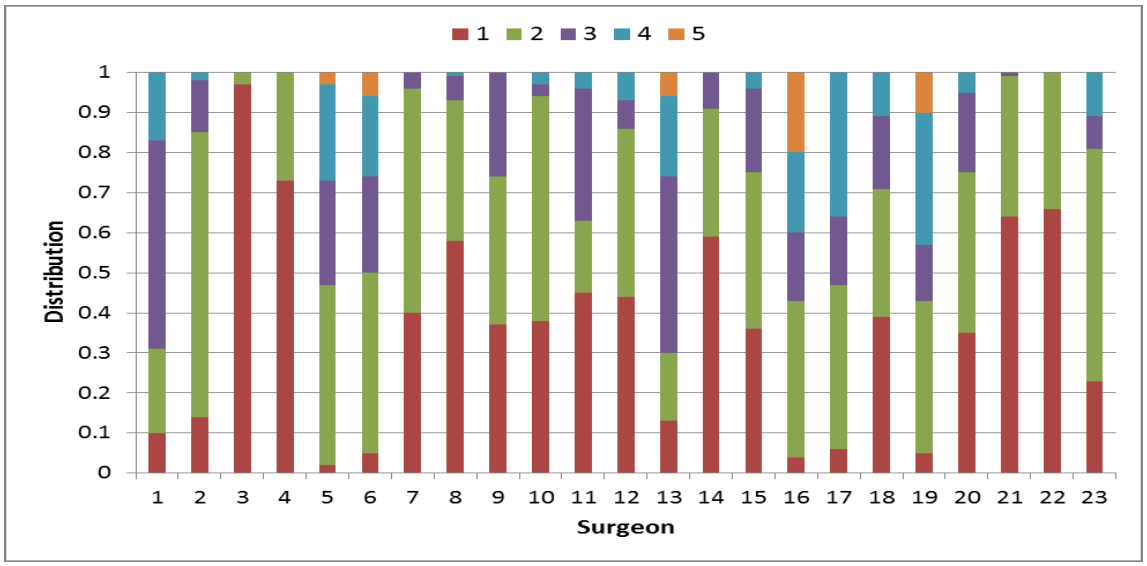

Figure 4: Inpatients per block by surgeons

Furthermore, we suspected that differences in each surgeon's practice and patient monitoring preferences could be a source of variability in patient length of stay even within the same specialty and for similar procedures. Therefore, we categorized surgical specialties based on CMGs and a co-morbidity index used by the hospital to identify the types of patients serviced by each surgeon. Figure 5 compares three surgeons` practice in discharging patients for one specific CMG (320) and for different co-morbidity index levels. Each color represents one surgeon. These surgeons deal with different population sizes for this specific group of patients. We verified the premise of different preferences of surgeons in discharging patients with similar treatment requirements using F-tests and T-tests for both variance and mean of the LOS distributions of corresponding surgeons. In all the three compared pairs of surgeons the null hypotheses for equal variances and equal means cannot be rejected.

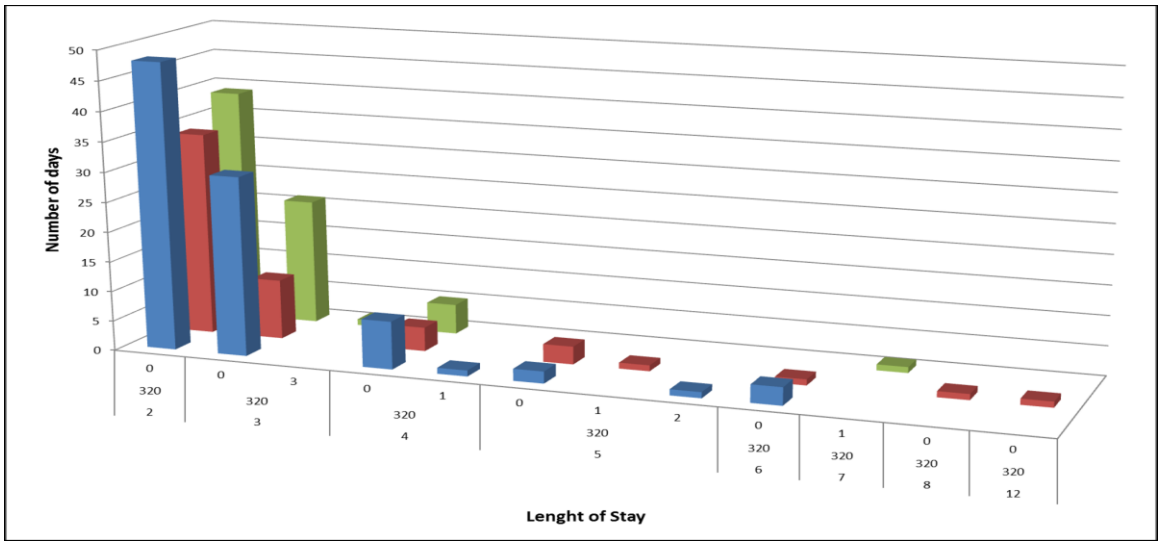

Figure 5: Surgeon` behavior in discharging patients with similar type of surgery

Even for cases such as shown in Figure 6 for CMG 502, where schematic evaluation of LOS distribution for different surgeons may suggest different discharge practices, the same statistical tests for variance and mean of LOS distributions per surgeon suggest no statistically significant difference. Nonetheless, because surgeons treat different types of patients, the distribution of the length of stay of each surgeon is quite different. Hence the importance of considering ward occupancy in designing the block schedule.

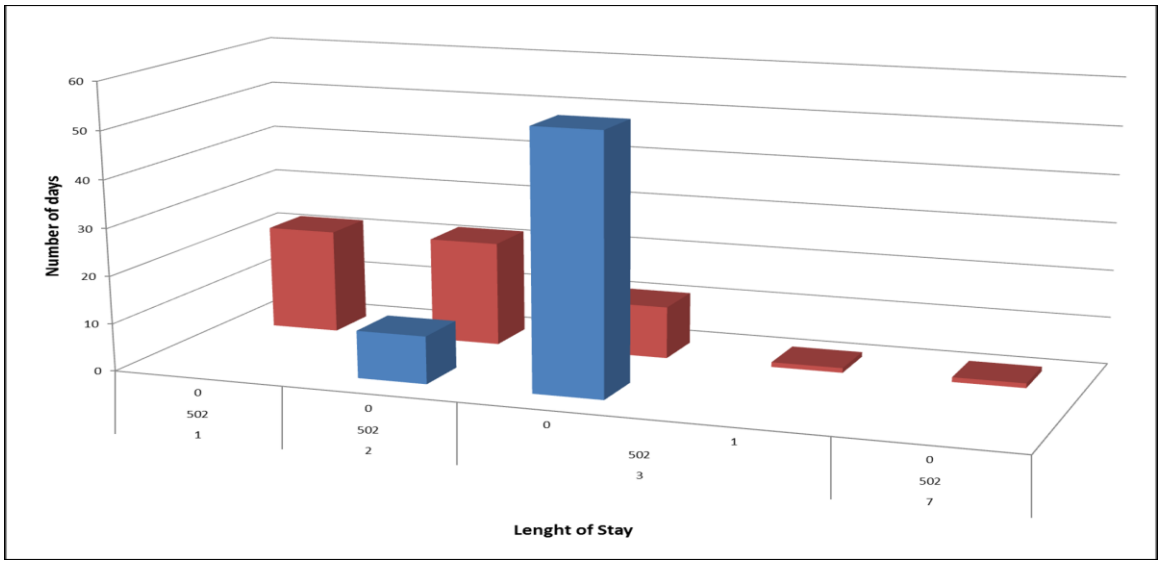

Figure 6: Surgeon` behavior in discharging patients with similar type of surgery

As explained above, each surgeon may receive a variety of patients requiring different types of surgery and with differing severity (co-morbidity). According to the above findings, we used the aggregated inpatient groups with different CMGs and co-morbidity levels under one surgeon in order to determine the LOS distribution. By considering the probability distribution for length of stay of patients for each surgeon, we can allow the allocation of surgical time to each surgeon in our model to depend on the stochastic nature of recovery time in the ward. It is important to note that this also allows us to create the block schedule on the basis of information available to hospital management. At the time the block schedule is created, the hospital can have no idea as to what type of surgeries the surgeon will book into a block. Thus, the only option available is to consider the distribution of LOS for each surgeon rather than delving below the surgeon level to each CMG.

The important co-efficient in the objective function of our model that we need to consider is $p_i^{LOS}(s)$ which represents the probability that a patient for surgeon i stays s days in the hospital. According to the historical data, we considered a discrete distribution for patient LOS for each surgeon. For instance, for surgeon A, patients may be discharged after one day with a probability of 20%, after two days with the probability of 40% and after three days with the probability of 40%; while surgeon B`s patients would be discharged with a 60% chance after one day and 40% chance after two days. Each surgeons` inpatient's length of stay follows a multinomial distribution. We assumed that no patient stayed longer than 28 days in the surgical ward – an assumptions that was easily verified by the data.

In addition, we were given the length of the master surgery schedule (4 weeks), the required number of blocks per cycle per surgeon and the number of available operating rooms.

Methodology

To address our research questions we applied Mixed Integer Programming (MIP) and later Robust Optimization (RO) to analytically model surgical scheduling. We built a flexible model that allocates OR capacity and provides scheduling guidelines in order to help improve QCH's ward utilization and to guard against the potential for high fluctuations in inpatient volume.

We verified and validated our analytical models and results using a simulation model. Furthermore, simulation provides us with the chance to test different scenarios, examine potential impacts on other parts of the hospital of the optimized master schedule and compare with the current situation in the hospital with revised schedules derived from the optimization model.

This section is organized in 3 major parts: In part A, we develop the MIP model in order to design a master surgery schedule for the hospital which results in a minimum overflow of patients from the surgical ward to the medical ward. We also extend our model to incorporate the surgeons` preferences for a consistent schedule each week and evaluate a number of scenarios with varying OR capacity on each day of the week. The performance measure used to evaluate each scenario is the expected peak bed occupancy at the ward over the length of a cycle. In part B, we use a simulation model of the medical and surgical wards to verify and validate the performance of solutions found in part A and to expand our scope to other interacting units and patient streams. In part C, we mathematically extend our model to a RO model to find an operational solution for situations where the hospital faces higher than average number of surgical inpatients. We

demonstrate the impact of these incidents of high congestion on the optimal surgical schedule.

A) Analytical Modeling: Integer Programming

Using mixed integer programming as a sub-category of mathematical modeling, programmers try to define a set of possible actions that can be taken, define the relationships and consequences of each action numerically and then try to optimize (whether minimize or maximize) an intended objective. Mixed Integer Linear Programming is a type of program in which the relationships of consequences of different actions are defined in a linear fashion and where decision variables can be continuous and/or integer. Actions are constrained by limitations on the availability of resources as well as other potential restrictions.

This methodology has been widely used in healthcare and hospital management. Whether or not to perform a surgery, allocate a bed, or allocate certain blocks to specific surgical specialties are all examples of binary decisions a hospital manager faces that have quantifiable consequences. Average number of required beds, overtime required or cost of a scenario are examples of continuous decision variables in the healthcare context.

Like any other modeling approach, the first step is to define different decision variables that are involved in the model. These variables represent the actions by which we can optimize our model. The second and third steps include defining constraints on the decision variables and setting up the objective function of the model. By the objective function we differentiate between what is favorable for the hospital and what is not, while constraints limit the set of possible actions. Our models include non-negative continuous variables and binary variables.

Mixed Integer Programming

Hospitals always try to maintain a set level of bed utilization with under-utilization and over-utilization both being seen as negatives. Ideally, they look to maximize throughput without generating bottlenecks and blockages in the wards. Maximizing throughput does not mean fully utilizing the available bed capacity. Instead, hospital managers maintain performance and throughput at a level below the total available beds in order to manage fluctuations in demand. A sudden surge in emergency arrivals or unusually long patient stays represent uncertain capacity requirements that can be covered by unutilized beds.

Since currently QCH struggles with overloads from the surgical ward to the medical ward, we set our objective to minimize the peak load in the surgical ward. The model is flexible enough to be used either to schedule surgeons or surgical specialties to specific OR blocks. Focusing at the surgeon level is based on the assumption of correlation between patients' LOS with a specific surgeon due to the type and mix of patients a given surgeon treats.

In this part, we initially follow Beliën & Demeulemeester (Beliën & Demeulemeester , 2007). However, we modify it later in order to include other properties according to QCH preferences.

To represent our model, we consider our decision variable as

$$X_{it} = \begin{cases} 1, \text{if physician } i \text{ is given an OR on day } t \\ 0, \text{otherwise.} \end{cases}$$

The model seeks to minimize the average peak load on the surgical ward over the scheduling period. Therefore, the objective function is as below:

$$Min_X Max_{t\in\{1,\dots,T\}} \left\{ \sum_{i\in[I]} \sum_{j\in\{1,\dots,T\},t\geq j} \sum_{s=t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) \right. \quad (1)$$

$$\left. + \sum_{i\in[I]} \sum_{j\in\{1,\dots,T\},t<j} \sum_{s=T+t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) \right\}$$

Where

T = The number of days in the master schedule,

$p_i^{LOS}(s) =$ The probability that a patient of surgeon i will have a length of stay of s days,

I = The number of physicians; and,

n_i = Average number of inpatients surgeon i sends to the ward per block.

The model assumes that the master schedule is repeated after T days. Based on this cyclic assumption, the objective function calculates the expected patient census for each day t (t between 1 and T) and chooses the block schedule (X) that minimizes the largest of these values (called the peak load). The expected patient census is calculated by looking back over the last T days and determining the expected number of patients still in the hospital by day t who received surgery on any of the T days leading up to day t. The first term in the objective deals with days in the same cycle that are prior to the current day t and the second term deals with days $t+1$ through day T from the previous cycle. The assumption is that no patient stays in the surgical ward longer than T days. Figure 7 helps depict this calculation.

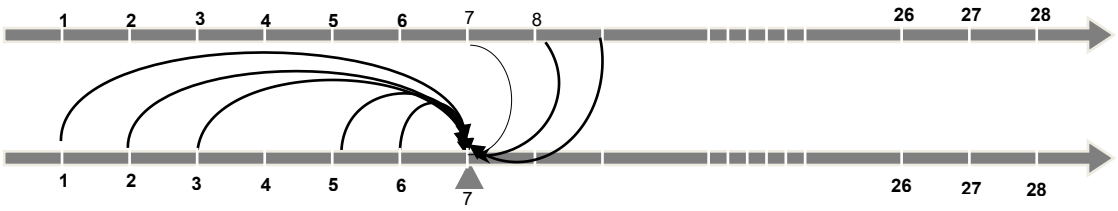

Figure 7: Elective patients' presence in ward at any day t from surgeries during last T days

Constraints for our model can be grouped into two main categories: constraints on the demand side and constraints on the supply side. On the demand side, we have both surgical demand which results in OR utilization and post-operative demand for beds. In our model, the demand for beds is incorporated in the objective function. On the other hand, we assume the demand for surgical capacity is equal to the current capacity allocated to each surgeon per month thus maintaining the current throughput. This constraint can be represented as:

$$\sum_{t\in[T]} X_{it} = d_i \qquad \forall i \in [I] \tag{2}$$

where

d_i = The number of blocks required for each surgeon i over T day period.

On the supply side, elective surgeries take place in the Main Operating Rooms. Surgeries for different specialties take place in 8 operating rooms the output of which forms the bed demand in the ward. Therefore, in order to manage bed utilization in the ward, we may need to change our policy for scheduling the OR. The Supply constraint on OR availability is represented as:

$$\sum_{i\in[I]} X_{it} \leq N_t \qquad \forall t \in [T] \tag{3}$$

where

N_t= The number of ORs available on day t of the master schedule.

This constraint ensures that we do not schedule more blocks into a day than available operating rooms.

Since the objective function is not linear, we reformulate the model as below and solve it to minimize a new decision variable μ which substitutes for the average maximum occupancy over T days.

$$Min_{X,\mu}` \mu \tag{4}$$

Subject to:

$$\sum_{i\in[I]} \sum_{j\in\{1,\dots,T\},t\geq j} \sum_{s=t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) + \sum_{i\in[I]} \sum_{j\in\{1,\dots,T\},t<j} \sum_{s=T+t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) \leq \mu \tag{5}$$

$$; \forall t \in [T]$$

$$\sum_{t\in[T]} X_{it} = d_i \qquad \forall i \in [I] \tag{6}$$

$$\sum_{i\in[I]} X_{it} \leq N_t \qquad \forall t \in [T] \tag{7}$$

In this formulation, constraint (5) calculates the expected patient census for each day t and forces it to be less than the maximum occupancy μ.

Results:

The model was programmed in Cplex solver using IBM-ILOG Optimization Studio IDE version 12.6. Two main scenarios were initially investigated. The first scenario follows the current practice of the hospital of opening a maximum of 8 ORs for 5 days a week. The second scenario considers the possibility of 2 open ORs on each day of the weekend. In both scenarios, the number of blocks required by each surgeon is held constant.

Resulting schedules for both scenarios demonstrate a significant reduction in the maximum peak load in a month when compared to the current master schedule. Moreover, the standard deviation in bed occupancy reduces dramatically especially for the scenario with weekend access. The results for bed occupancy in the surgical ward are represented in Table 1:

Criteria	Current Schedule	Only Weekdays	Including Weekends
Average Max	33.24	26.99	24.82
Average Min	10.02	13.03	21.70
Monthly Ave	23.69	23.69	23.69
STD	7.64	4.32	0.79

Table 1: Results of MIP model for the current schedule and the two scenarios

As the above results suggest, the MIP model can provide the hospital with a schedule which leads to a lower peak load in the ward. Moreover, it provides a higher minimum

occupancy, and so, reduces the deviation thus providing a more predictable load. This means that the new schedule provides the ward with a smoother occupancy level during a cycle. The new master schedule reduces peak load level by about 19% even when restricted to the same 5 days a week. When 2 ORs are opened on each day of the weekend the results are even more significant with a further drop of more than 2 beds in the peak load and an even greater reduction in the standard deviation.

Considering the need to staff for the potential peak load and considering that an additional nurse is required for every 4 patients on the ward, a reduction in peak load of 8 patients results in 2 fewer full time nurses required per shift.

So far, according to the objective function, we are targeting bed occupancy at the ward and modify the OR schedule in order to maintain the improvement in bed peak load. However, we need to consider the consequences of such alterations on the utilization of ORs. A new scheduling scenario affects the OR utilization pattern over the cycle which may have its own advantages and disadvantages for the hospital.

Opening ORs during the weekends not only smooths out the bed utilization in the ward, but also smooths out OR utilization over the weekdays. Figure 8 represents the difference in OR utilization between the optimum schedule of the scenario with 2 ORs per day during the weekends compared to the optimized schedule considering only weekdays.

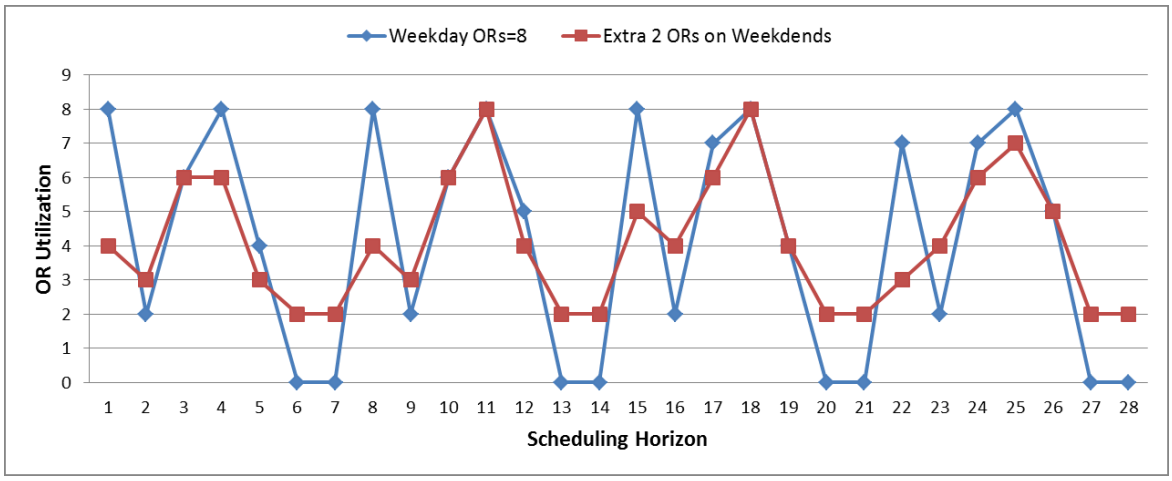

Figure 8: OR utilization change by opening 2 ORs per day in weekends

Therefore, although expanding the functioning of the operating rooms to the weekends requires surgeons and other OR staff to be available; it requires less of the same resources during weekdays. Also, with only two ORs per day on the weekends, it would be possible to enhance the model to ensure a rotation of surgeons through the weekend blocks so that no surgeon is required to work on the weekends more than once a month. Nonetheless, hospital management needs to evaluate the costs and benefits of such a drastic change on all stakeholders such as staff, nurses and surgeons.

Model including hard constraints on surgeons` availability

Due to surgeons` commitments at other hospitals and clinics, the master surgery schedule needs to be as consistent as possible from week to week for each surgeon. In order to incorporate this constraint, we define a new decision variable and add some hard constraints as detailed below.

The new decision variable Y_{iw} takes the value 1 if surgeon i has been allocated a block on day w of the week and takes the value 0 otherwise. The intent is to allocate to a given surgeon only as many weekdays as required to meet his/her demand for blocks. Thus a

surgeon who needs 4 or fewer blocks in a month should receive them all on the same day of the week while a surgeon who needs between 5 and 8 blocks in a month should receive them on two days of the week.

The general formulation below represents the limitation we imposed on allocating same day of the week throughout the master schedule to each surgeon. For each surgeon, constraints for seven days of the week are considered. We are using the big M method to prevent allocation on other days of the week to a surgeon once a sufficient number of days have been allocated. Constraint (8) ensures that the new decision variable is forced to equal one if surgeon *i* receives a block on any week day w in the cycle:

$$X_{it} + X_{i(t+7)} + X_{i(t+14)} + X_{i(t+21)} \leq M\, Y_{iw} \quad ; i \in [I], t \in [T], w \in [W] \tag{8}$$

Constraint 9 ensures that a given surgeon is allocated blocks on a minimum number of weekdays.

$$\sum_{w} Y_{iw} \leq \left\lceil \frac{d_i}{4} \right\rceil \tag{9}$$

Results:

Similar scenarios with and without weekends were run including the hard constraints and the subsequent schedules were obtained. The results for occupancy are given in Table 2:

Criteria	Current Schedule	Without Weekends	With Weekends
Average Max	33.24	27.09	24.78
Average Min	10.02	13.97	21.75
STD	7.64	4.15	0.77

Table 2: Results of IP model including hard constraints

As can be seen, there is little reduction in the added value of the resulting schedules compared to the case where no constraints on weekly consistency were included. This is largely because the original model without constraints on consistency nevertheless produced a master surgery schedule that did provide reasonable consistency. The reason stems from the varying lengths of stay of patients of different surgeons that results in a weekly pattern in allocation of blocks for the master surgery schedule cycle.

Scenario development and analysis

In our model development and data analysis, we assumed that incorporating surgeon effect on LOS can affect the surgical schedule and consequently the occupancy in the ward. To verify this assumption, according to our second research question, we try one scenario of our model using the same LOS distribution for all surgeons. The distribution is derived by aggregating the data of LOS of all patients for all surgeons. As the result, the average maximum occupancy at ward would be 31.79 beds, while by incorporating different LOS distribution for each surgeon; this level would be 27.09 beds thus demonstrating the importance of incorporating the variability in LOS.

To better evaluate the scenarios discussed earlier, we varied the number of available ORs both on weekday and weekends. The results are presented below:

Scenario:		Average Max	Average Min	STD
# of ORs per weekday	# of ORs for weekends			
6	0	27.83	13.31	4.72
6	2	24.76	22.10	0.72
7	0	27.21	11.56	4.61
7	2	24.88	18.67	1.19
10	0	26.95	13.35	4.14
10	2	24.79	20.93	0.88
8	1	25.33	18.64	1.81
8	3	24.62	22.15	0.74
8	2 only on Saturdays	25.69	18.26	2.12

Table 3: Different scenario evaluation

Comparing the results presented in Table 3 to those provided in Table 2, opening extra ORs during weekdays does not greatly improve the maximum occupancy, the minimum occupancy or STD. The minimum required ORs for weekdays is 6 ORs per day in order to maintain a feasible constraint on total OR availability. By adding extra ORs for weekdays, the peak load and occupancy deviation at ward do not improve much. However, the main improvement can be caused by opening ORs during weekends. Figures 9 and 10 demonstrate the effect of adding OR availability to the weekdays and weekends.

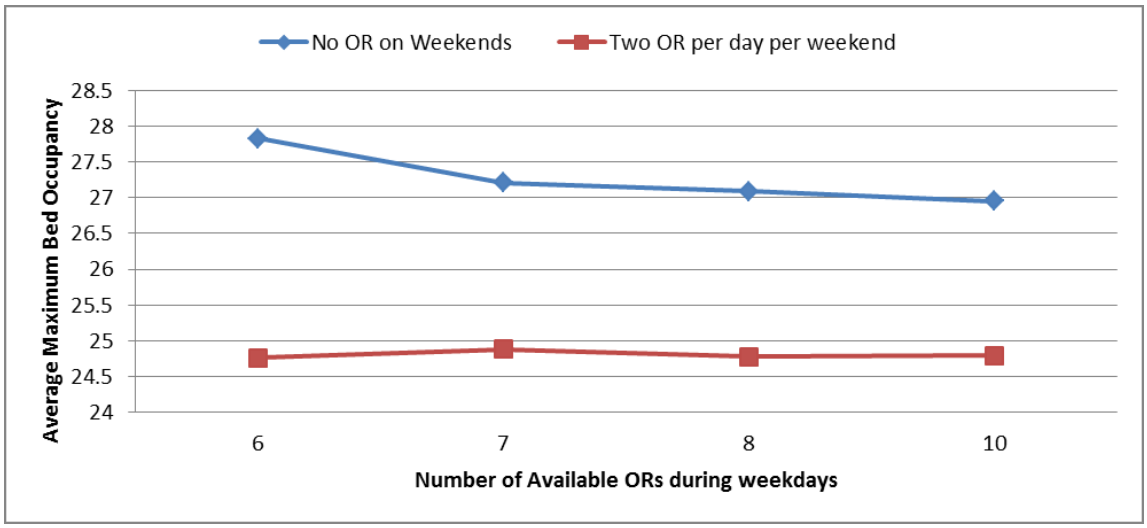

Figure 9: Effects of adding ORs to weekdays and weekends on maximum peak load at ward

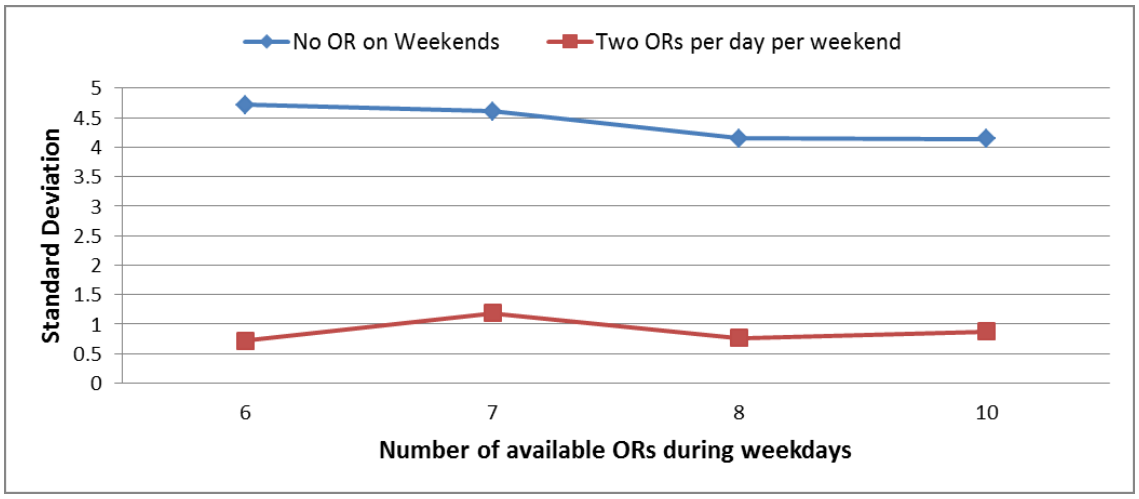

Figure 10: Effects of adding ORs to weekdays and weekends on occupancy variation at ward

As the above figures suggest, providing extra ORs on weekdays does not offer much improvement in bed occupancy while maintaining the same OR availability for weekends; the main improvement in bed occupancy is due to the addition of ORs on the weekends. However, different availability of ORs per weekday can noticeably affect the OR utilization distribution over a cycle.

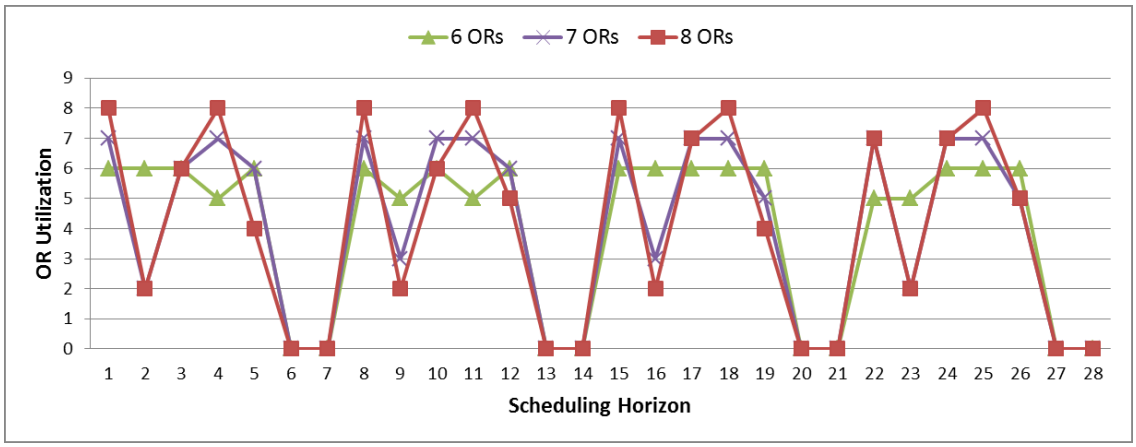

Figure 11: OR utilization distribution for different availability scenario

According to Figure 11 extra OR availability during weekdays causes more variation in OR utilization in the surgical unit over the cycle. By imposing limits on OR availability, the hospital can maintain a smoother utilization over the cycle. However, it highly depends on hospital policies since we are only scheduling for elective cases that require a post-surgery stay in the hospital whereas the hospital deals with other patient streams such as emergency and day surgeries that may require additional ORs to be opened.

It is worth nothing that the maximum OR allocation per day does not change by opening more than 8 ORs per weekday as the maximum allocated ORs during the planning cycle remains 8.

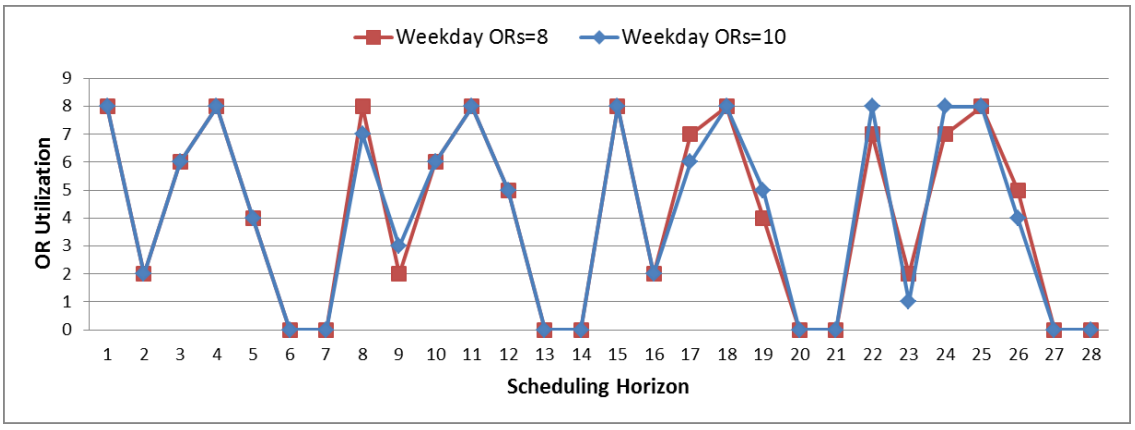

Figure 12: OR utilization Distribution of scenarios 8 and 10 OR availability per weekday

However, as Figure 12 suggests, the master schedule is slightly different as we have arrived at an equally good (based on the objective function) but slightly different OR allocation based on the expanded feasibility set. This implies that there are multiple optimal solutions for this problem. Adding more than 8 ORs provides the model with additional extreme points, so in order to find the optimal solution; the solution algorithm takes a different path to find an optimal solution. This leads to a different but equally optimal solution.

On the other hand, by adding extra capacity for weekends, we see a relative improvement in the objective function although with a diminishing marginal effect. As seen in Figure 13, by adding extra OR time during weekends even up to 3 OR per day, the solution suggests using all available ORs during the weekends. Although changes in bed utilization distribution is not significant, open ORs during weekends has a significant effect on the OR utilization distribution over the horizon.

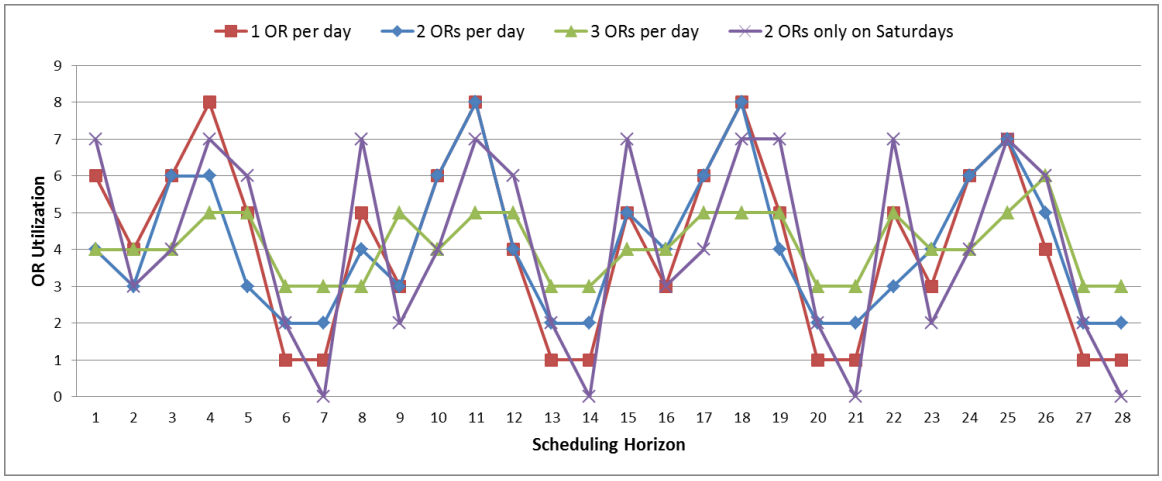

Figure 13: OR utilization for different number of operating ORs during weekends

In this section, we used a linear model which reflects resource limitations in the surgical unit including ward and ORs in order to improve bed occupancy level in the ward. Later we added surgeons` preference to maintain similar individual weekly. We could provide the hospital with a new schedule which offers lower bed requirement at peak load times.

Moreover, we tested some intuitive scenarios and their possible consequences on performance measures such as maximum peak load in the ward and OR. Our results generally suggest that opening extra ORs during weekends may have a significant effect on performance measures but only up to a point (after which additional OR time on the weekends has little impact) while extra ORs on weekdays do not offer any improvement.

Although we included the variability of LOS of patients for each surgeon, our models so far were based on average inpatient referrals per surgeon per block. In part C, we develop a robust model to include uncertainty in the number of generated inpatients per block by each surgeon.

To validate and verify our models and findings, we developed a simulation model which is explained in the following section. Our simulation model addresses a wider scope other than just the surgical department in the hospital.

B) Simulation- Validation and Verification

Models and simulations are important tools for analyzing systems. Models are mathematical constructs that describe the performance of subsystems. Interactions among subsystems in a larger system, combined with the constraints within which the system operates, influence the performance of the total system. Using these models and simulations, it becomes possible to analyze the expected performance of a system if systemic changes are made. In this section, we develop a simulation model with a wider scope than what we addressed in the previous mathematical models. In the mathematical models, we concentrated on the bed occupancy of the surgical ward resulting from elective surgeries. In the simulation, we cover both the medical and surgical ward. Moreover, we include medical patients as well as elective and emergency patients.

Simulation provides us with the chance to further monitor the effect of changes made in the master surgery schedule on the wider hospital environment.

Simulation has been widely applied in the healthcare literature. Jun et al (Jun, J. B, Jacobson, S. H, & Swisher, J. R, 1999) provide a review of over 100 papers concerning simulation in healthcare. A common use of simulation in hospitals is to model the effect of process changes on system behavior. From a healthcare application perspective, Jun et al (Jun, J. B, Jacobson, S. H, & Swisher, J. R, 1999) define Discrete Event Simulation (DES) as an operational research technique that allows the end user to assess the efficiency of existing healthcare delivery systems, to ask 'what if?' questions regarding potential changes and to design new systems. DES allows stakeholders to assess the efficiency of a healthcare system and test the applicability of different scenarios for re-design. In DES, events such as patient arrivals to the surgical ward or patient discharges from the ward take place at discrete points in time. DES also allows the user to incorporate significant detail about the system into the model by simulating each individual patient with unique properties. For example, patients can arrive at different times, have different length of stay, have different ailments and require different resources.

A key aspect of a DES model is the system state description that includes values for all of the variables in the system. If any variable changes, it changes the system state. In a simulation, the dynamic behavior of the system can be observed as entities (e.g., patients) move through the nodes and activities (e.g., OR, surgical ward, medical ward) identified in the model. Often a warm up period to reach a steady state, or else, an initial state of the system can be inputted. The model is then tested to see if it describes the performance of

the existing system. Once the model has been validated, it can be used to explore the consequences of different actions and predict the future.

We generated our simulation of the system using Arena software for 13 elective patient categories, 10 emergency categories and 30 medical patients categories. The classification of elective cases in the simulation model is different from the previous mathematical model. In the mathematical model, all patients were aggregated under the attending surgeon independent of their CMGs or co-morbidities. However, in the simulation model, elective, emergency and medical cases were categorized based on CMG and co-morbidity as the two factors most likely to best predict LOS. Based on historical data, the distribution of patient types that each surgeon refers to the ward is determined and the patient type is used to determine the length of stay distribution. Different co-morbidities under one CMG were aggregated in order to gain a reasonable population size for each patient type.

The other freedom that simulation provides us is the possibility of incorporating different monthly block schedules whereas in the mathematical model we could consider only one repeating schedule. This is helpful in terms of mimicking current practice where there are differences in the schedule from one month to the next.

Poisson distributions were fitted to emergency and medical patients' arrivals. As Figures 14 and 15 depict, for aggregated emergency CMGs and medical CMGs, patients' arrivals depict similar behavior to Poisson distributions to an acceptable extent.

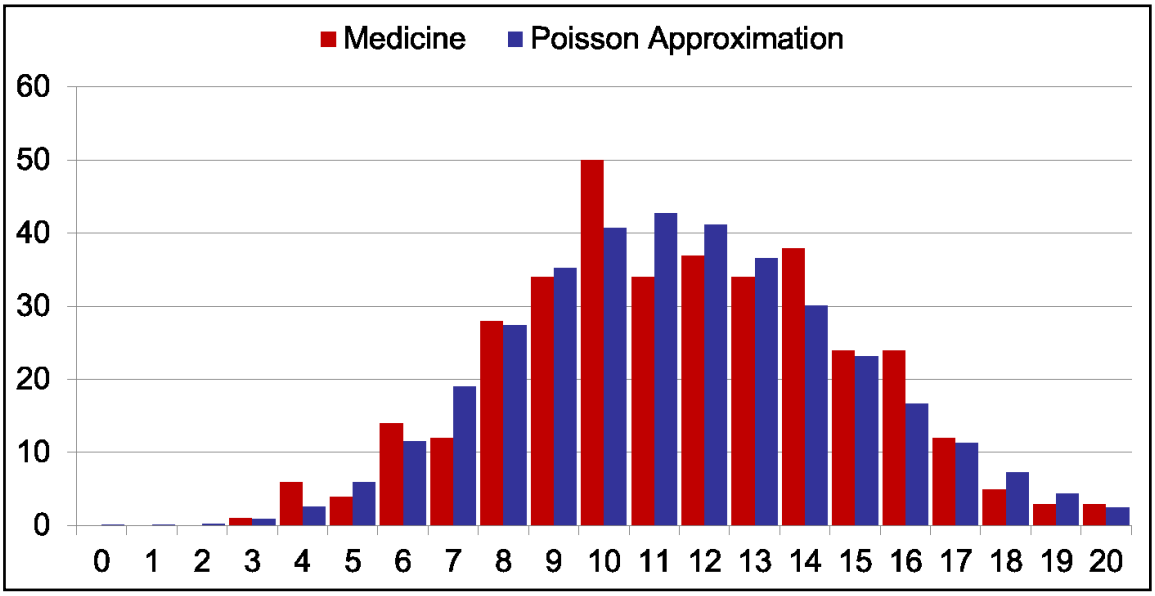

Figure 14: Arrival Distribution for Medicine

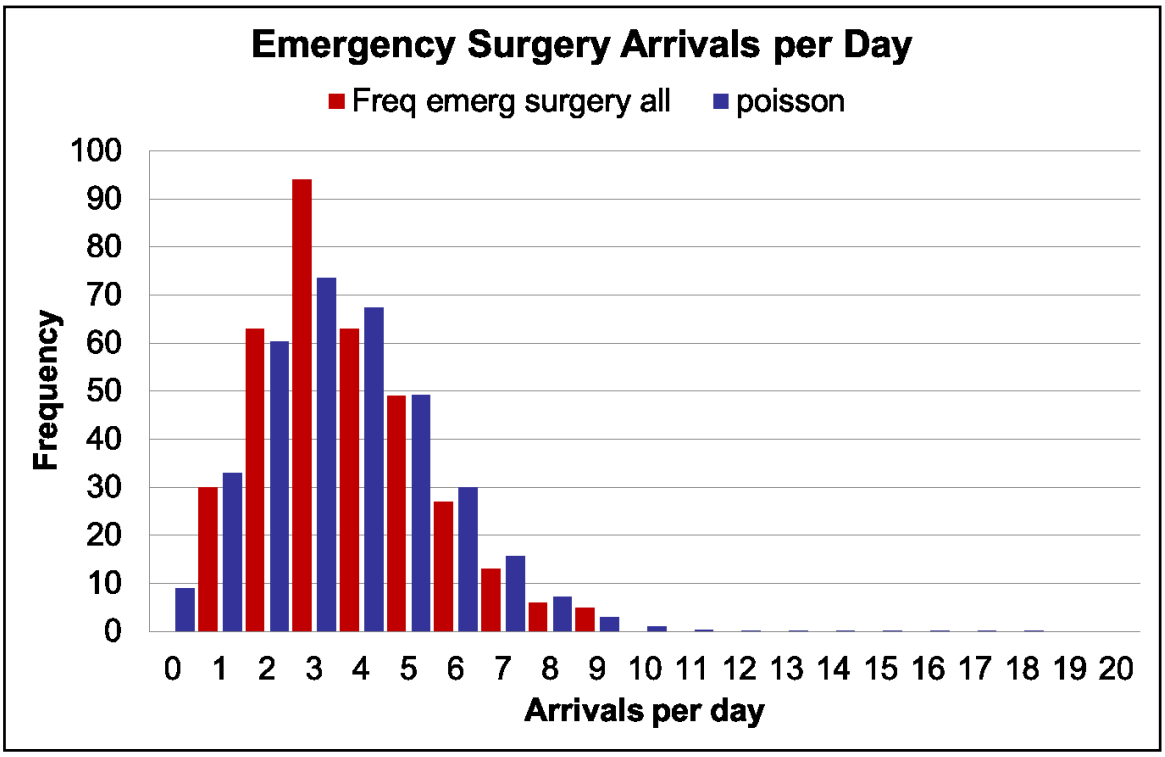

Figure 15: Arrival distribution for Emergency

We validated our simulation model according to the bed occupancy level at both the medical and surgical wards. An initial state was fed into the model based on the census on January 1st, 2013 and the simulation was run for one year. Census for that year was then compared to the simulated census. Figures 16 through 18 represent a comparison of

our simulated distributions of hospital census and the actual data for medical patients as well as emergency and elective surgical patients.

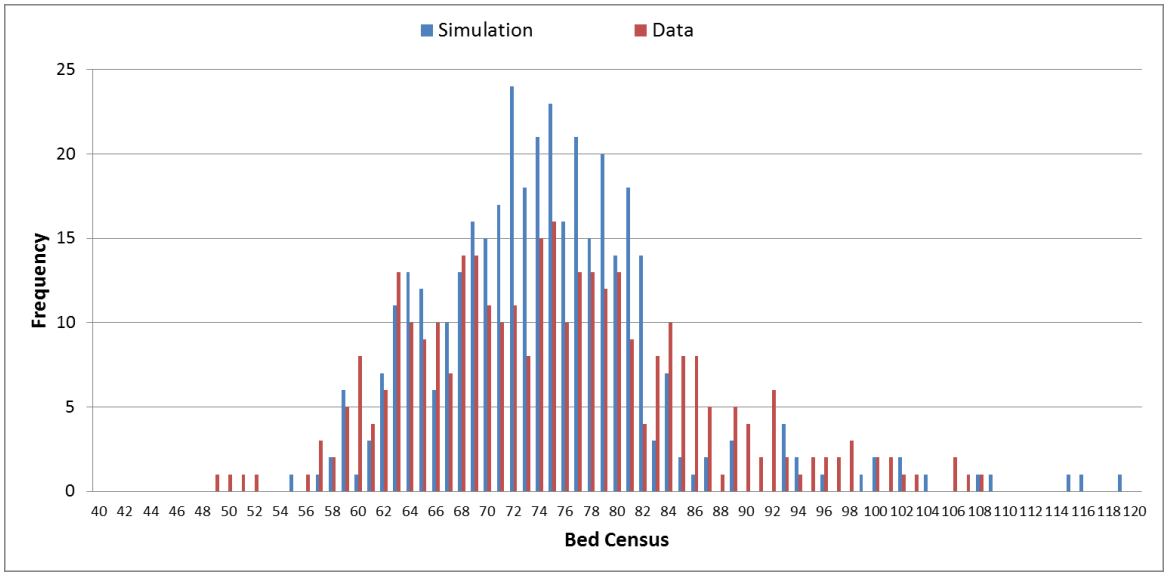

Figure 16: Medical patients` census counts

Figure 16 depicts the census for medical patients. For both census output of data and simulation, average census is about 75 beds with similar distributions and standard deviations (≈4.5 for simulation output and 6 for actual data).

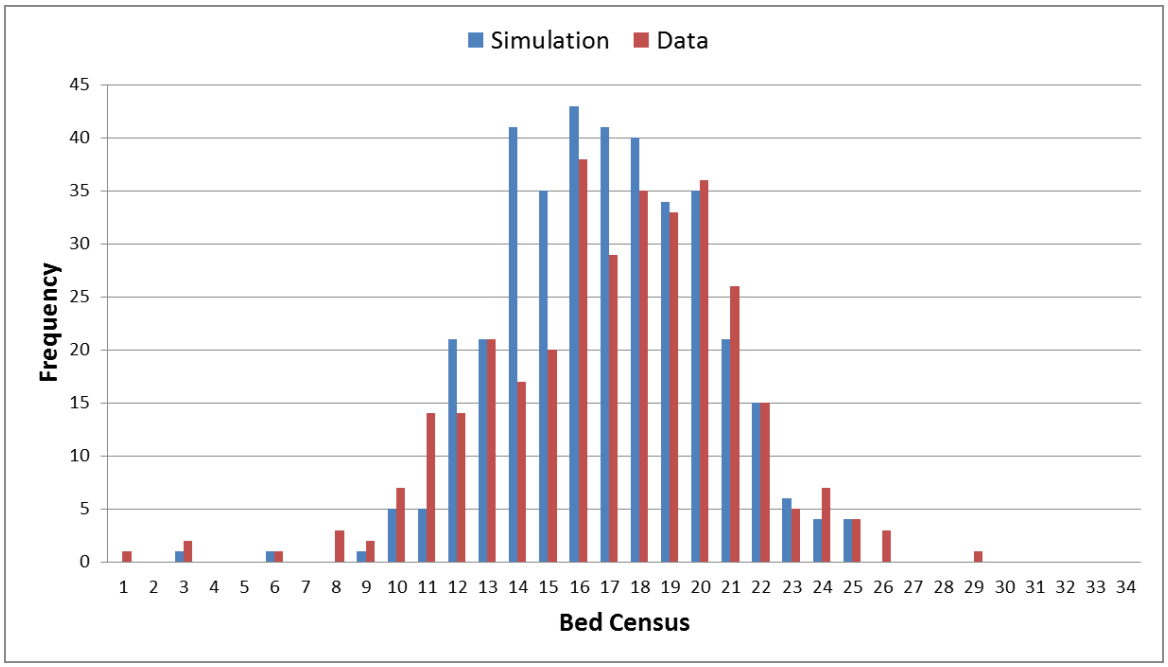

Figure 17: Emergency surgery census count

Simulated emergency surgery census also follows the distribution of actual emergency surgery census at ward. Both averages are about 17 beds and standard deviation for simulation is about 9.5 versus 8 for actual data.

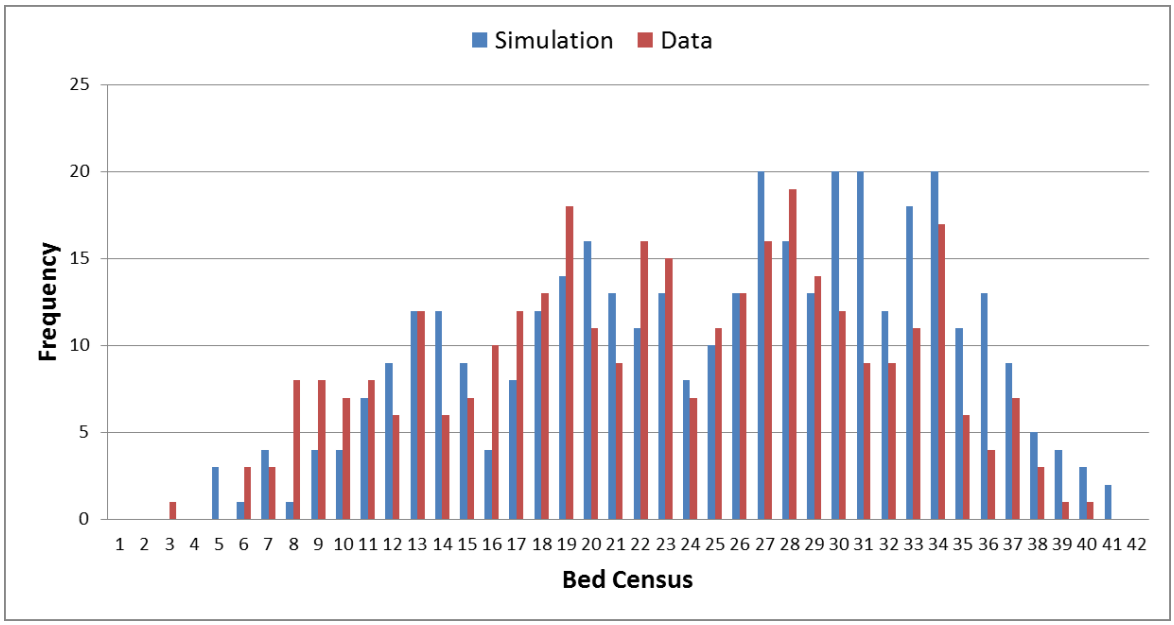

Figure 18: Elective Surgery census count

For elective bed occupancy (Figure 18), the simulated distribution average and standard deviation are 25 and 5.6 respectively, whereas data represents values of 23 and 5.

As seen in Figures 16 to 18, the simulation outputs for the bed census distribution for all the three major categories of patients mimic the actual census distribution derived from the empirical data quite well.

Now that our simulation model is validated we can use it to validate our MIP model. Figure 19 represents the effect of the optimized schedule derived from the MIP model for both the scenario with only weekday ORs and the scenario with two ORs on each day of the weekends as compared to the current schedule.

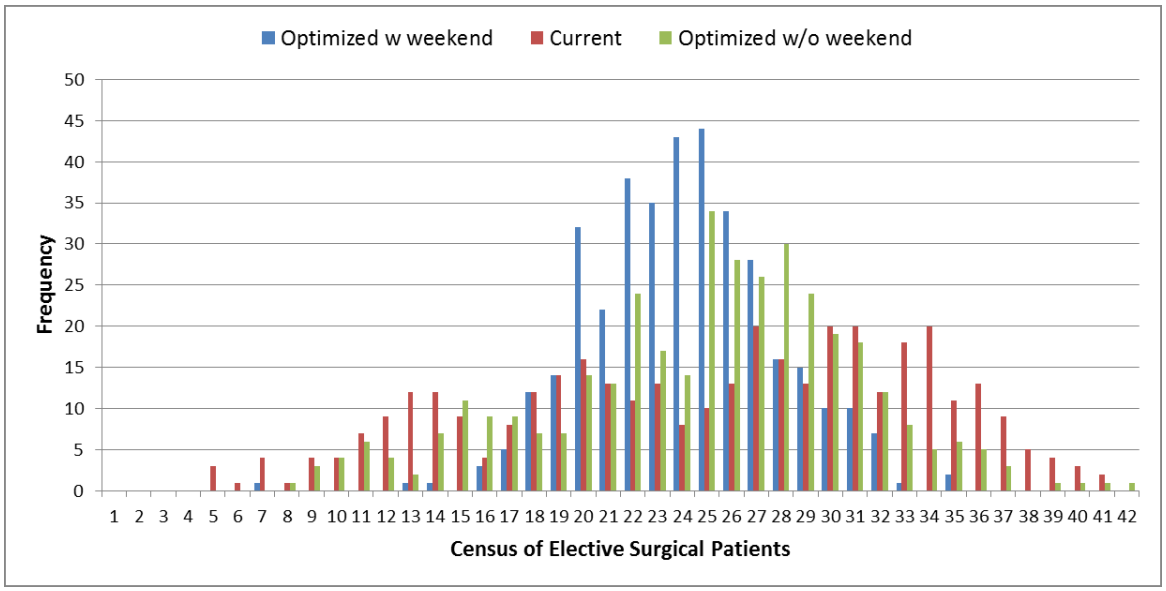

Figure 19: Simulated bed census using optimized schedules

As the comparison suggests the maximum occupancy and occupancy variation in the ward are reduced significantly by the optimized schedule – as suggested by our analytical model. Figure 20 shows how the occupancy variation from day to day is reduced during the week.

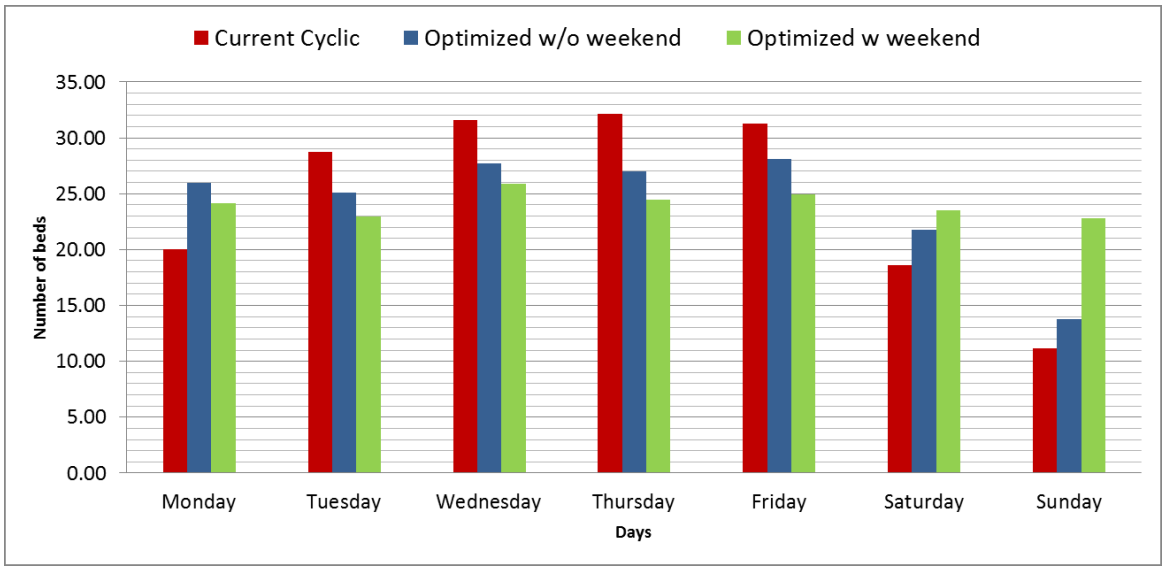

Figure 20: Comparison of Average Week-day Census of Elective Patients

According to Figure 20, the new schedules decrease the occupancy in the ward for Tuesdays, Wednesdays, Thursdays and Fridays while increasing the occupancy for

Mondays, Saturdays and Sundays. It can be noted that opening two ORs during weekends has a significant effect in variation reduction.

C) Robust Optimization

As mentioned earlier, the number of inpatients referred to the ward by each surgeon from each block is a stochastic variable. Using average values as the basis for our model mandates that surgeons try to maintain the number of scheduled procedures per block close to the average in order to ensure our expected outcomes in the ward. In other words, surgeons need to be encouraged to spread out their inpatients evenly between their blocks. However, this imposed behavior may be in conflict with what is the best action for patients. Therefore, we need to extend our model in such a way as to incorporate some uncertainty with respect to the patient load in a given block.

The master surgery schedule is designed for a 4-week cycle based on the average inpatient referral rate by each surgeon. However, we may face periods of high congestion due to some surgeons producing more than the average amount of inpatients in their blocks. Therefore, we need to design the master schedule in such a way as to accommodate different ranges of demand, or at least, taking into account the potential for periods of high congestion. In this section, we apply Robust Optimization to find the best solution for periods of high congestion.

Robust optimization is a method that helps manage uncertainties in different parameters of the model including cost coefficients of the objective function or coefficients in the right hand side of the constraints. The main paradigm in RO is worst-case analysis by which a solution is evaluated according to the most unfavorable realization (Gabrel, Murat , & Thiele , 2014) of a subset of the uncertain parameters. RO tries to provide

protection from adverse realizations that result in a trade-off with system performance. For RO modeling, we need to define a realistic set of values of the uncertain parameter which is called the "uncertainty set". RO tries to optimize the objective function over all possible scenarios according to the uncertainty set. RO models may experience complexity issues due to the inclusion of larger sets of decision variables and constraints. The choice of uncertainty sets may play an important role in the level of complexity and computational tractability. Uncertainty sets may be defined as polyhedral or ellipsoidal. Polyhedral uncertainty sets lead to linear robust programming while ellipsoidal sets will have a second-order conic program counterpart (Gabrel, Murat , & Thiele , 2014). One focus of RO literature is to develop methods that resolve such complexities (see (Ben-Tal, El Ghaoui, & Nemirovski, 2009) and (Bertsimas & Sim, 2003)). In these studies, duality theory has been heavily used in order to reformulate a RO model into a computationally tractable problem. Another issue associated with the use of RO is over-conservatism. In order to control over conservatism according to practical purposes we limit the uncertainty set based on historical values of patient referrals for each surgeon. We follow the work of (Bertsimas & Sim, 2003) and apply their design of the uncertainty set in modeling the uncertainty associated with the number of inpatients referred to the ward per block.

In part A, we used probability distributions to accommodate the stochasticity of the patients' length of stay in our model. However, our data analysis, as represented in Figure 3, shows that surgeons may refer a variable number of inpatients during each surgical block as well whereas the model presented in part A is based on the average number of inpatients per surgeon per block. We use RO to accommodate this aspect of uncertainty in bed occupancy. To optimize the model under uncertainty, we can use either stochastic programming in which parameters follow a known distribution; or we can use robust

optimization in which parameters are known within some bounds. According to the structure of our formulation, we used stochastic distribution of patients` length of stay for each surgeon. However, using probability distribution of number of referrals per surgeon per block to the ward in our model construct results in estimated value of number of referrals which is again average referral (n_i).

In particular, by RO, we allow n_i to be stochastic within a range of average referral level and maximum referral level. We seek to minimize the worst case scenario for the peak load when a subset of surgeons refer more than the average number of inpatients for the ward in a day. We used the minimax approach according to the framework introduced by (Bertsimas & Sim, 2003) to minimize the worst case loss in the objective value that may occur.

Stochastic n_i per surgeon

To incorporate the stochasticity in the per block referral of inpatients by each surgeon into our model we consider the possibility of each surgeon exceeding his average inpatient referral from every given block.

Recall that our initial objective function is:

$$Min_X Max_{t\in\{1,\dots,T\}} \left\{ \sum_{i\in[I]} \sum_{j\in\{1,\dots,T\},t\geq j} \left[\sum_{s=t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) \right] \right. \tag{10}$$

$$\left. + \sum_{i\in[I]} \sum_{j\in\{1,\dots,T\},t<j} \left[\sum_{s=T+t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) \right] \right\}$$

If we define $F(X_{ij})$ as:

$$F(X_{ij}) = Max_{t\in\{1,\dots,T\}}\left\{\sum_{i\in[I]}\sum\nolimits_{j\in\{1,\dots,T\},t\geq j}[\sum_{s=t-j+1}^{T} X_{ij}n_i p_i^{LOS}(s)]\right. \quad (11)$$

$$\left. + \sum_{i\in[I]}\sum_{j\in\{1,\dots,T\},t<j}[\sum_{s=T+t-j+1}^{T} X_{ij}n_i p_i^{LOS}(s)]\right\}$$

Then, the initial objective function could be written as:

$$Min_X F(\mathrm{X_{ij}}) \quad (12)$$

Therefore, to maintain the initial objective of minimizing the peak load in the ward in the robust framework, we define ε_i as the proportion of deviation from the average towards the maximum inpatient referral by each surgeon i and δ_i as the maximum increase above the average in the number of inpatients per block for surgeon i. Our objective function for minimizing the maximum bed occupancy at ward would then be:

$$Min_X Max_\varepsilon F(\mathrm{X_{ij}}, \varepsilon_i) \quad (13)$$

That would be expanded as:

$$Min_X Max_{\varepsilon_i\in\left\{\substack{0\leq\varepsilon_i\leq 1\\ \sum_i \varepsilon_i\leq\gamma}\right\}} Max_{\ t\in\{1,..,T\}}\left\{\sum_{i\in[I]}\sum_{j\in\{1,..,T\},t\geq j}\left[\sum_{s=t-j+1}^{T} X_{ij}(n_i\right.\right. \quad (14)$$

$$\left. + \varepsilon_i.\delta_i)p_i^{LOS}(s)\right]$$

$$\left. + \sum_{i\in[I]}\sum_{j\in\{1,..,T\},t<j}\left[\sum_{s=T+t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i)p_i^{LOS}(s)\right]\right\}$$

This formulation of the objective function minimizes the peak load in the ward as before, but now with the additional complexity that some surgeons may exceed their average inpatient load in a block. $Max_{\varepsilon_i \in \begin{Bmatrix} 0 \le \varepsilon_i \le 1 \\ \sum_i \varepsilon_i \le \gamma \end{Bmatrix}}$ in this formula offers the most adversarial deviations from the average inpatient referral by surgeons during the cycle conditioned on the total deviation being less than the control parameter γ . It helps with generating the worst-case scenario according to the possibility of higher referral by a subset of surgeons. As such, the formulation computes and minimizes the worst-case (maximum) peak load resulting from the potential deviations. Each surgeon has the chance to deviate from the average inpatient referral to the ward with the potential variability incorporated using the terms ε_i and δ_i. ε_i belongs to the uncertainty set which represents the percentage of deviation from the average inpatient referral per day per surgeon. Since ε_i is a continuous decision variable, the robust model is a Mixed Integer problem. γ is the degree of freedom that the system owner can consider for each day on deviations from the set of average inpatient referral. We discuss later how each ε_i can only take the extreme values of the corresponding constraint (0 or 1). Due to this, γ tells us the maximum number of surgeons permitted to refer inpatients at the maximum level on each specific day *t*. By this formulation we choose the best schedule that minimizes the worst case scenario of bed occupancy resulting from uncertainties in the number of inpatient referrals and length of stay over every scheduling day *t*.

The constraints on the above objective function are as before:

$$\sum_{t \in [T]} X_{it} = d_i \qquad \forall i \in [I] \tag{15}$$

$$\sum_{i\in[I]} X_{it} \le N_t \qquad \forall t \in [T] \tag{16}$$

$$X_{it} + X_{i(t+7)} + X_{i(t+14)} + X_{i(t+21)} \le M\, Y_{iw} \qquad ; i \in [I], t \in [T], w \in [W] \tag{17}$$

$$\sum_w Y_{iw} \le \left\lceil \frac{d_i}{4} \right\rceil \tag{18}$$

To solve the model, we take the following steps to reformulate it so that that Equation (14) becomes computationally tractable and can be solved by off-the-shelf solvers. In order to accomplish this, we need to linearize the model. As before, we introduce the dummy variable μ which represents the count of bed occupancy on each day *t* and bounds the worst-case peak load from above. Therefore, our objective function seeks to minimize μ and the model would be:

$$Min_{X,\mu,\varepsilon}\ \mu \tag{19}$$

Subject to:

$$Max_{\varepsilon_i \in \left\{ \substack{0\le\varepsilon_i\le 1 \\ \sum_i \varepsilon_i \le \gamma,} \right\}} Max_{t\in\{1,..,T\}} \left\{ \sum_{i\in[I]} \sum_{j\in\{1,..,T\},t\ge j} \left[\sum_{s=t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i) p_i^{LOS}(s) \right] \right. \tag{20}$$

$$\left. + \sum_{i\in[I]} \sum_{j\in\{1,..,T\},t<j} \left[\sum_{s=T+t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i) p_i^{LOS}(s) \right] \right\} \le \mu$$

$$\sum_{t\in[T]} X_{it} = d_i \qquad \forall i \in [I] \tag{21}$$

$$\sum_{i\in[I]} X_{it} \le N_t \qquad \forall t \in [T] \tag{22}$$

$$X_{it} + X_{i(t+7)} + X_{i(t+14)} + X_{i(t+21)} \leq M\, Y_{iw} \quad ; i \in [I], t \in [T], w \in [W] \tag{23}$$

$$\sum_w Y_{iw} \leq \left\lceil \frac{d_i}{4} \right\rceil \tag{24}$$

Mathematically, we can switch the order of the two maximization expression in constraint (20) that leads to the following formulation:

$$Max_{t\in\{1,..,T\}} Max_{\varepsilon_i \in \left\{ \substack{0\leq\varepsilon_i\leq 1 \\ \sum_i \varepsilon_i \leq \gamma,} \right\}} \left\{ \sum_{i\in[I]} \sum_{j\in\{1,..,T\}, t\geq j} \left[\sum_{s=t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i) p_i^{LOS}(s) \right] \right. \tag{25}$$

$$\left. + \sum_{i\in[I]} \sum_{j\in\{1,..,T\}, t<j} \left[\sum_{s=T+t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i) p_i^{LOS}(s) \right] \right\} \leq \mu$$

Constraint (25) represents the maximum occupancy over T days of the cycle resulting from the maximum inpatients generated in all the days j before day t within the limit of T days ($T \geq s$). To linearize the model we first re-write constraint (25) as below:

$$Max_{\varepsilon_i \in \left\{ \substack{0\leq\varepsilon_i\leq 1 \\ \sum_i \varepsilon_i \leq \gamma} \right\}} \left\{ \sum_{i\in[I]} \sum_{j\in\{1,..,T\}, t\geq j} \left[\sum_{s=t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i) p_i^{LOS}(s) \right] \right. \tag{26}$$

$$\left. + \sum_{i\in[I]} \sum_{j\in\{1,..,T\}, t<j} \left[\sum_{s=T+t-j+1}^{T} X_{ij}(n_i + \varepsilon_i.\delta_i) p_i^{LOS}(s) \right] \right\} \leq \mu$$

$$; \forall t \in [T]$$

The left-hand side of Constraint (26) captures the worst-case occupancy for each day t= $1, \ldots T$ under the most adversarial choice of ε_i.

From this point on, we demonstrate how to deal with the maximization problem for each constraint. We first separate the terms including the decision variables ε_i. The purpose is to separate the terms that make the model non-linear so that in later steps, we can address those terms and modify them in order to reach a linear model. Equation (27) is another representation of constraint (26):

$$\sum_{I\in[I]}\sum_{j\in\{1,\dots,T\},t\geq j}\sum_{s=t-j+1}^{T} X_{ij}n_i p_i^{LOS}(s) \tag{27}$$

$$+\sum_{i\in[I]}\sum_{j\in\{1,\dots,T\},t<j}\sum_{s=T+t-j+1}^{T} X_{ij}n_i p_i^{LOS}(s)$$

$$+\ Max_{\varepsilon_i\in\left\{\substack{0\leq\varepsilon_i\leq 1\\ \sum_i\varepsilon_i\leq\gamma}\right\}}\left\{\sum_{i\in[I]}\sum_{j\in\{1,..,t\}}\left[\sum_{s=t-j+1}^{T} X_{ij}\varepsilon_i\delta_i p_i^{LOS}(s)\right]\right.$$

$$\left.+\sum_{i\in[I]}\sum_{j\in\{t+1,..,T\}}\left[\sum_{s=T+t-j+1}^{T} X_{ij}\varepsilon_i\delta_i p_i^{LOS}(s)\right]\right\}\leq\mu$$

$$;\forall t\ \in[T]$$

In this constraint, the first two terms are those seen previously in part A of the Methodology section and are linear whereas the last two terms are non-linear due to the product of two decision variables X_{ij} and ε_i. According to (Bertsimas & Sim, 2003) we can solve this model by linearizing the non-linear part. Therefore, we define $\beta^t(\gamma, X_{ij})$ for each $t \in [T]$ as:

$$\beta^t(\gamma, X_{ij}) = Max_{\varepsilon_i}\left\{\sum_{j\in\{1,..,t\}}\left[\sum_{s=t-j+1}^{T} X_{ij}\varepsilon_i\delta_i p_i^{LOS}(s)\right]\right. \tag{28}$$

$$\left. + \sum_{j\in\{t+1,..,T\}}\left[\sum_{s=T+t-j+1}^{T} X_{ij}\varepsilon_i\delta_i p_i^{LOS}(s)\right]\right\}$$

Subject to:

$$0 \le \varepsilon_i \le 1; \quad \forall i \in [I] \tag{29}$$

$$\sum_i \varepsilon_i \le \gamma \tag{30}$$

Note that the objective function in (28) is dependent on the day *t*, but we would apply the same set of constraints for every day.

The maximization problem (28)-(30) is in fact a special case of the Continuous Knapsack Problem. In the continuous knapsack problem, the objective is to fill a knapsack with fractional amounts of different items in order to maximize the used volume. It is proven that either items are not chosen to be included in the knapsack or the item is chosen to its full amount. Hence, the optimal ε_i can be guaranteed to take either values 0 or 1 when γ is integer.

According to the strong duality property of a linear program with finite optimal value, $\beta^t(\gamma, X_{ij})$ has a dual form with equal finite optimal value. $\beta^t(\gamma, X_{ij})$ is a maximization model with a bounded feasible region and thus meets the condition for duality by holding finite optimal points. Therefore, if y_i corresponds to each $\varepsilon_i \le 1$ constraint and z corresponds to the constraint $\sum_i \varepsilon_i \le \gamma$, the dual form of $\beta^t(\gamma, X_{ij})$ would be:

$$Min \ \sum_i y_i^t + \gamma z^t \tag{31}$$

Subject to:

$$y_i^t + z^t \geq \sum_{j\in\{1,..,t\}} \left(\sum_{s=t-j+1}^{T} X_{ij}\delta_i p_i^{LOS}(s) \right) + \sum_{j\in\{t+1,..,T\}} \left(\sum_{s=T+t-j+1}^{T} X_{ij}\delta_i p_i^{LOS}(s) \right); \quad \forall i \in [I] \tag{32}$$

$$y_i^t, z^t \geq 0 \tag{33}$$

By replacing $\beta^t(\gamma, X_{ij})$ in the main expression (27) for each day *t*, we arrive at the following:

$$Min_{X,Y,y,z,\mu} \ \mu \tag{34}$$

Subject to:

$$\sum_{i\in[I]} \sum_{j\in\{1,\dots,T\}, t\geq j} \sum_{s=t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) + \sum_{i\in[I]} \sum_{j\in\{1,\dots,T\}, t<j} \sum_{s=T+t-j+1}^{T} X_{ij} n_i p_i^{LOS}(s) + \sum_i y_i^t + \gamma z^t \leq \mu \ ; \ \forall t \in [T] \tag{35}$$

$$y_i^t + z^t \geq \sum_{j\in\{1,..,t\}} \left(\sum_{s=t-j+1}^{T} X_{ij}\delta_i p_i^{LOS}(s) \right) + \sum_{j\in\{t+1,..,T\}} \left(\sum_{s=T+t-j+1}^{T} X_{ij}\delta_i p_i^{LOS}(s) \right);\ \forall i \in [I]\ , \forall t \in [T] \quad (36)$$

$$\sum_{t\in[T]} X_{it} = d_i \qquad ;\ \forall i \in [I] \quad (37)$$

$$\sum_{i\in[I]} X_{it} \leq N_t \qquad \forall t \in [T] \quad (38)$$

$$X_{it} + X_{i(t+7)} + X_{i(t+14)} + X_{i(t+21)} \leq M\ Y_{iw} \qquad ; i \in [I], t \in [T], w \in [W] \quad (39)$$

$$\sum_{w} Y_{iw} \leq \left\lceil \frac{d_i}{4} \right\rceil \quad (40)$$

$$y_i^t, z^t \geq 0 \quad (41)$$

The above model was coded in Cplex. We verified the model by comparing its result with the result of the initial model based on (Beliën & Demeulemeester , 2007) by setting $\gamma = 0$. Then we investigate the effect of increasing γ on the schedule and consequently the bed occupancy levels in the ward. It is worth noting that, as constraint (35) depicts, in this model we are considering the same γ for all days *t*.

Once we solve the model, we are able to extract more information from the result. In particular, in what follows we explain how we are able to identify which day is the worst-case day and which surgeons are most critical in each scenario for different γ values. We take the following steps to identify the peak load day and critical surgeons:

1) Identify the worst-case day representing the highest peak load:

Suppose that X^*, Y^*, Z^* are the optimal solutions of the problem corresponding to X_{ij}, y_i^t and z^t variables . We can expect that by replacing these matrixes in constraint (35) we can calculate the worst-case occupancy for each day *t* and choose the maximum as the worst-case maximum peak load over the cycle; the corresponding day would be the worst-case day. However, in our experiment, we observe that the solutions generated by Cplex do not provide us with such an easy solution due to degeneracy issues.

Degeneracy here refers to the presence of multiple optimal solutions that lead to the same optimal value. While the solution generated by Cplex is optimal in the sense that it solves the problem (34), the solution is actually not optimal to subproblem (31) for all t (which is not necessary for solving (34)).

Therefore, in order to find the optimal Y^* and Z^*, we solve the dual form of $\beta^t(\gamma, X_{ij})$ conditioned to X^*. Using the results we are able to calculate and identify the worst-case day.

2) Identify the critical surgeons who contribute to the worst case day in the robust model:

The right-hand side of constraint (36) represents the effect of deviation by any surgeon on each day on peak load (refer to definition of $\beta^t(\gamma, X_{ij})$ and its dual form). We can simply calculate this expression for each surgeon for the day with the worst peak load conditioned on X^*. The γ largest values corresponding to surgeons can indicate the γ critical surgeons for the corresponding scenario. Since for each t, z^t is the same, this could be equivalent to finding γ largest $y_i$for every t. For example, for $\gamma = 1$, surgeon 8

is the one with the maximum right-hand side value of constraint (36) and for $\gamma = 2$, surgeons 21 and 8 represent the 2 largest values.

Results:

Table 4 shows the effect of increasing γ on the maximum occupancy level in the ward for the situation in which we have 2 open ORs for each day of the weekend and 8 open each week day. In this table we show which doctors on which days are those who refer the maximum possible inpatients to the ward.

γ	Maximum Occupancy	Surgeons selected to max-out	Blocks per month	Average Inpatient referral	Maximum referral
0	24.494	-	-	-	-
1	26.919	#8	4	1.5	4
2	28.571	#21	5	1.37	3
		#8	4	1.5	4

Table 4: Effect of increasing stochasticity in surgeons` inpatient referral level on surgical unit performance

Table 4 suggests that by increasing the degree of freedom for surgeons to refer more inpatients than the average the ward occupancy level increases. According to our results, surgeons who are selected to maximize the deviation from the average are not necessarily those with the highest possible inpatient load. The surgeon with the highest possible inpatient load is surgeon #5 with 7 blocks per month and a maximum of 5 inpatients per block.

Figure 21 represents the change of OR utilization with new schedules as γ is increased:

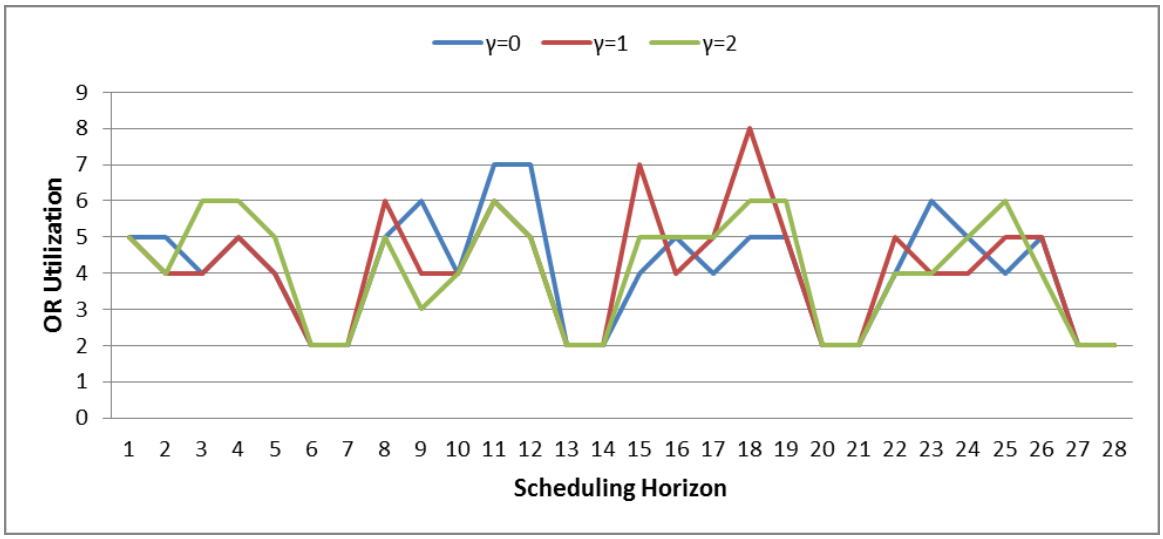

Figure 21: OR utilization for Robust schedules according to ε_i

This Figure suggests that there is in fact a significant shift in the optimal master schedule as we attempt to guard against the worst case scenario.

In addition, we compared the schedules produced by the robust models against the original optimal solution and under the assumption that all surgeons refer inpatients at the average rate. This allows us to see the "cost" associated with guarding against the worst case scenario in terms of average performance. Table 5 provides the expected peak load for the various models:

	$\gamma = 1$	$\gamma = 2$
Average referral rate in robust schedule	25.03	24.58
Initial Optimum Schedule	24.49	

Table 5: Effect of using average number when some surgeons refer more than average inpatients to the ward

This suggests that guarding against minor deviations (1 or 2 surgeons) can be done with little deterioration to the average performance of the system.

We were also interested to determine how much improvement the robust schedule can provide compared to a nominal schedule optimized based on the average of inpatient referrals. We evaluate both schedules under the worst-case scenario that there are γmany

surgeons who refer more than average inpatients and the surgeons considered are the ones generating the worst possible peak load. This helps us determine the impact of planning for periods of congestion as opposed to simply planning for the average. Table 6, compares the maximum occupancy at ward resulting from each schedule for γs equal 1 and 2:

	$\gamma = 1$	$\gamma = 2$
Schedule for average referral	28.27	29.51
Robust schedule	26.92	28.57

Table 6: Effect of using different schedule when deviating from average inpatient referral

Table 6, shows how using average numbers can be misleading in bed management, if, in reality, we face stochasticity in the number of procedures in each block in scheduling ORs. If hospital managers ignore the effect of uncertainty in the number of referred inpatients to the ward per block per surgeon in OR scheduling, they may face unexpected periods of congestion in the ward.

Comparison between the results of Table 5 and Table 6 suggest that the "cost" in performance in the average case when following the robust schedule is significantly less than the "cost" in performance in the congested scenario when following the optimal schedule based on averages.

Scenario development and analysis

Unfortunately, running the robust model is somewhat computationally burdensome. Thus, in order to investigate the effect of a larger number of deviating surgeons from the average inpatient referral on ward occupancy for new robust schedules, we need to decrease the size of the model.

In one scenario with the aim of scheduling for 14 days, we obtained the results in Table 7 for different values of γ. It is worth noting that for $\gamma = 0$ we obtained a relatively higher value compared to the original scenario due to the fact that we are running a shorter time horizon. We had to adjust the number of required blocks per cycle per surgeon and also the maximum length of stay of patients and thus the scenarios are not comparable. For example, since the cycle is halved, the required blocks per cycle is halved as well. However, the model only accepts the integer values so we decided to round up the non-integer halved values. These adjustments result in the presence of more patients in the ward.

	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 3$	$\gamma = 4$
Average Maximum Occupancy	24.67	27.153	29.013	30.91	32.69

Table 7: Occupancy level at ward with different Gammas

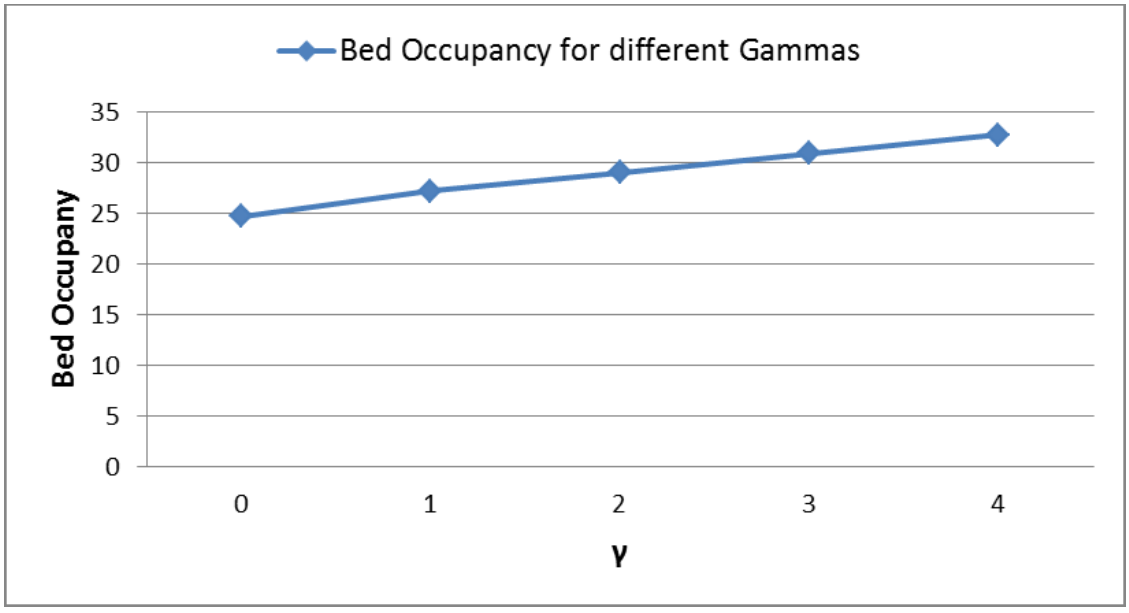

Figure 22: Bed Occupancy for different Gammas - case of 27 surgeons

To have more visibility on the effect of γ, we also considered a scenario with less surgeons. In this scenario, we reduced the number of surgeons from 27 surgeons to 10

without any specific selection criteria. We also reduced the cycle size from 28 to 14. However, we did not adjust the number of ORs available per weekday but increased the weekend OR availability to 8 as well. The results of the increasing level of freedom as γ increases for scheduling 10 surgeons for 14 days are shown in Figure 23.

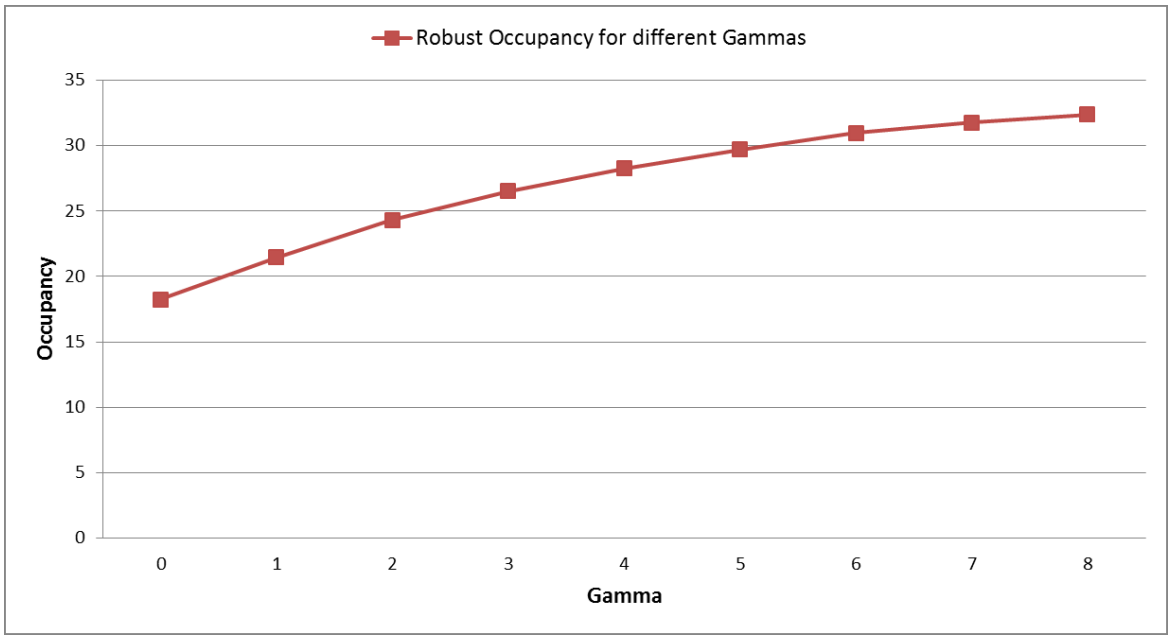

Figure 23: Effect of increasing Gamma on bed occupancy level at ward

As perceived from Figures 22 and 23, by increasing γ the occupancy increases but the marginal increase decreases. It should be noted that although we have 10 surgeons, we would not increase γ more than 8 since we have a maximum of 8 ORs available per day and so a maximum of 8 surgeons scheduled. Therefore, according to the definition of γ and the maximum value an ε_{ij} may receive, the greater degree of freedom (γ) cannot be utilized in the model and the constraint $\sum_i \varepsilon_{ij} \leq \gamma$ would be binding.

Figure 24 compares the occupancy resulting from using the robust schedule for different γs when each surgeon's referral is based on the average against the occupancy resulting from the optimal schedule for the average case.

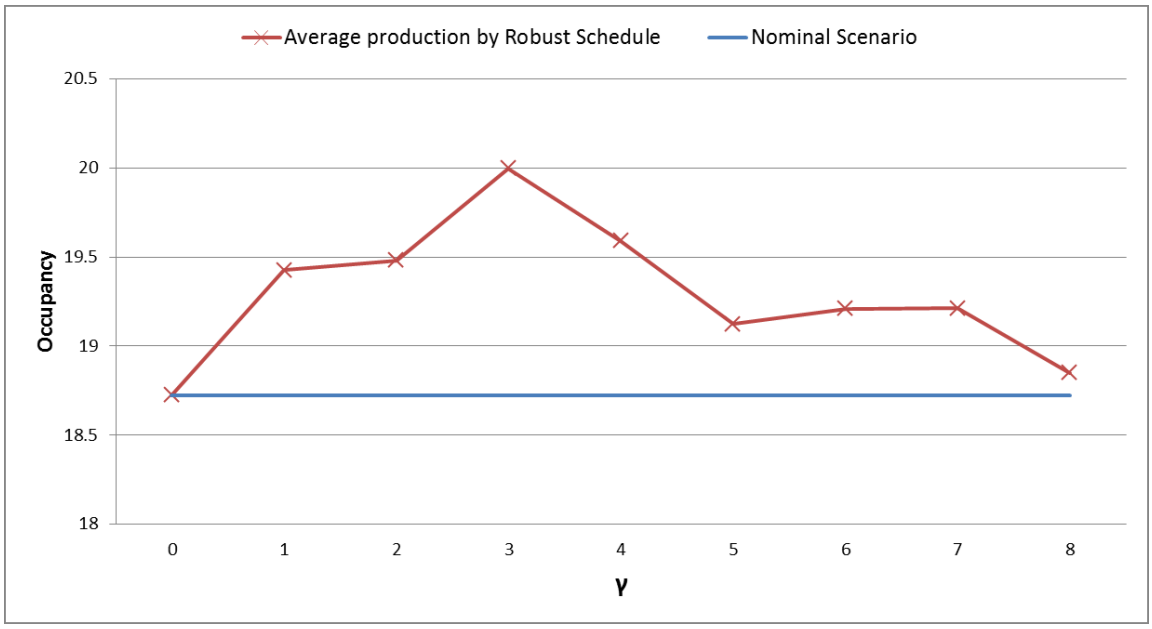

Figure 24: Average inpatient referral in robust schedules for different γs vs. occupancy in the nominal model

Figure 24 suggests that if in a robust schedule, surgeons refer the average number of inpatients to the ward; the ward would rarely face overflow. On the other hand, the issue of underutilization may arise if the hospital bases its actions according to the maximum deviation but surgeons follow the average inpatient referral level.

In another analysis, we investigated critical surgeons in different scenarios corresponding to different γs in order to determine the effect of deviation from the average by the same surgeons in the average schedule. The results for different γ are represented in Figure 25:

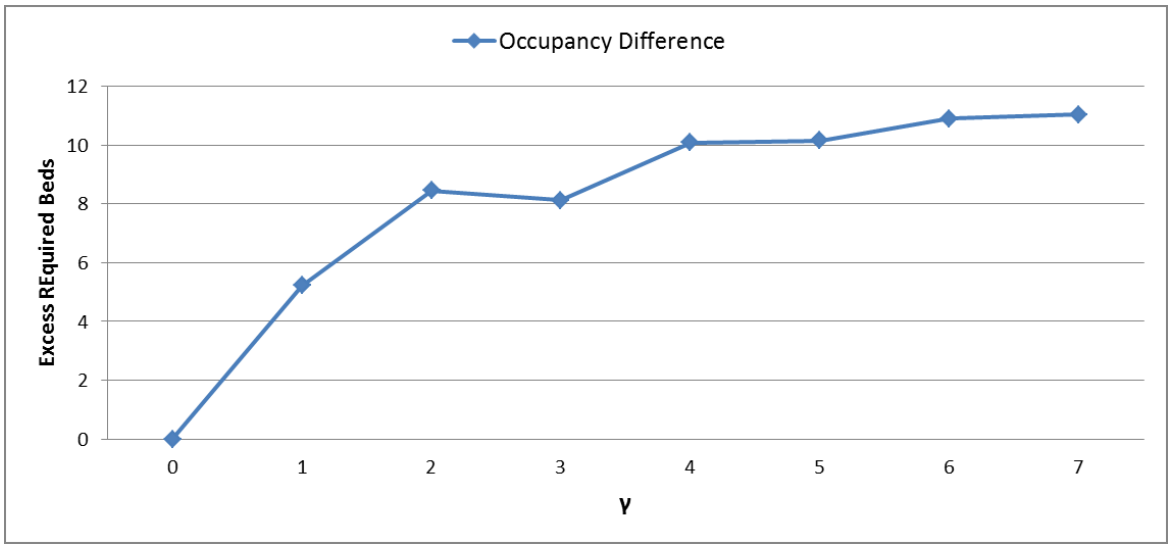

Figure 25: Comparison between occupancy in robust scenario vs. deviation in the nominal scenario

Figure 25 suggests the increase in peak loads resulted from the robust schedules based on increasing γs versus the peak occupancy resulting from the nominal schedule by deviating the corresponding surgeons from the average referral level. Moreover, as illustrated by Figure 25, the impact of increasing γ reduces as γ increases. This behavior could be explained from two perspectives: the problem setting and solutions dynamics. In our problem setting, the total OR requirement is 45 ORs per cycle while we made 112 ORs available. Average operating OR per day is 3 ORs and maximum OR utilization in the nominal scenario is 5. Therefore, even if γ increases substantially, it would be only used for a limited number of blocks per critical surgeon during the cycle. The other issue that may contribute to the decreasing marginal increase of peak loads for greater γs is the solution dynamics and the choice of critical surgeons. According to the structures of the objective function and uncertainty set, we can expect that more critical surgeons are chosen by smaller γs and that increasing γ results in the contribution of less critical surgeons, and therefore, a smaller increase in maximum occupancy.

Stochastic n_i per surgeon per block

By robust modeling of our problem, so far we considered the effect of increased inpatient referrals by surgeons. One limitation of the above model is that if any surgeons are recognized as the critical surgeons who may deviate according to different levels of γ, they would deviate to the maximum inpatient referral level in *all* their blocks over a cycle. However, we can extend our model to allow different surgeons to deviate on different days.

For this purpose, we would introduce ε_{ij} as the proportion of deviation from the average inpatient referral by surgeon i on day j. Therefore, the objective function for our model would be:

$$Min_X Max_{\varepsilon_{ij}} Max_{\ t\in\{1,..,T\}} \left\{ \sum_{i\in[I]} \sum_{j\in\{1,..,T\},t\geq j} \left[\sum_{s=t-j+1}^{T} X_{it}(n_i + \varepsilon_{ij}.\delta_i) P_i^{LOS}(s) \right] + \sum_{i\in[I]} \sum_{j\in\{1,..,T\},t<j} \left[\sum_{s=T+t-j+1}^{T} X_{it}(n_i + \varepsilon_{ij}.\delta_i) P_i^{LOS}(s) \right] \right\} \quad (42)$$

In defining the uncertainty set, we have the following flexibility to constrain it:

- We can consider the limitation on number of surgeons that would be allowed to deviate on each day

$$\sum_{i} \varepsilon_{ij} \leq \gamma_j, \quad \forall j \in [T] \quad (43)$$

where γ_j is the maximum level of freedom for deviation on each day and can vary for different days

- We can consider the limitation on number of blocks that each surgeon can deviate during one cycle.

$$\sum_{j} \varepsilon_{ij} \leq \theta_i, \quad \forall i \in [I] \tag{44}$$

where θ_i would be less than or equal to the total demand for a surgeon in a cycle

- We can also impose limitation on the total number blocks that are allowed to deviate above the average

$$\sum_{i,j} \varepsilon_{ij} \leq \omega, \quad \forall i \in [I], \forall j \in [T] \tag{45}$$

Therefore, the main model could be written as:

$$Min_X Max_{\varepsilon_{ij} \in E} Max_{\ t \in \{1,..,T\}} \left\{ \sum_{i \in [I]} \sum_{j \in \{1,..,T\}, t \geq j} \left[\sum_{s=t-j+1}^{T} X_{it}(n_i + \varepsilon_{ij}.\delta_i) P_i^{LOS}(s) \right] + \sum_{i \in [I]} \sum_{j \in \{1,..,T\}, t<j} \left[\sum_{s=T+t-j+1}^{T} X_{it}(n_i + \varepsilon_{ij}.\delta_i) P_i^{LOS}(s) \right] \right\} \tag{46}$$

Subject to:

$$\sum_{t \in [T]} X_{it} = d_i \qquad \forall i \in [I] \tag{47}$$

$$\sum_{i \in [I]} X_{it} \leq N_t \qquad \forall t \in [T] \tag{48}$$

$$X_{it} + X_{i(t+7)} + X_{i(t+14)} + X_{i(t+21)} \leq M\, Y_{iw} \quad ; i \in [I], t \in [T], w \in [W] \tag{49}$$

$$\sum_{w} Y_{iw} \leq \left\lceil \frac{d_i}{4} \right\rceil \tag{50}$$

Set E where $\varepsilon_{ij} \in E$ is defined by the following constraints:

$$\sum_{i} \varepsilon_{ij} \leq \gamma_j \quad \forall j \in [T] \tag{51}$$

$$\sum_{j} \varepsilon_{ij} \leq \theta_i \quad \forall i \in [I] \tag{52}$$

$$\sum_{\mathrm{i,j}} \varepsilon_{\mathrm{ij}} \leq \omega \quad \forall \mathrm{i} \in [\mathrm{I}], \forall \mathrm{j} \in [\mathrm{T}] \tag{53}$$

$$0 \leq \varepsilon_{ij} \leq 1 \quad \forall i \in [I], \forall j \in [T] \tag{54}$$

In the following, we prove that in this formulation ε_{ij} would take values of 0 or 1.

The matrix representation of the uncertainty set

$$\varepsilon_{ij} \in \begin{Bmatrix} 0 \leq \varepsilon_{ij} \leq 1 \quad \forall i \in [I], \forall j \in [T] \\ \sum_{\mathrm{i}} \varepsilon_{\mathrm{ij}} \leq \gamma_j \quad \forall \mathrm{j} \in [\mathrm{T}] \\ \sum_{\mathrm{j}} \varepsilon_{\mathrm{ij}} \leq \theta_i \quad \forall \mathrm{i} \in [\mathrm{I}] \\ \sum_{\mathrm{i,j}} \varepsilon_{\mathrm{ij}} \leq \omega \quad \forall \mathrm{i} \in [\mathrm{I}], \forall \mathrm{j} \in [\mathrm{T}] \end{Bmatrix} \tag{55}$$

Could be as $A\varepsilon \leq b$ if ;

$$A = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{bmatrix} \tag{56}$$

Transpose of ε is $[\varepsilon_{11} \quad \varepsilon_{12} \ldots \quad \varepsilon_{1j} \quad \varepsilon_{21} \varepsilon_{22} \cdots \quad \varepsilon_{2j} \cdots \quad \varepsilon_{IT}]$,

And transpose of b is $[1 \cdots 1 \quad \gamma_1 \cdots \gamma_{\mathrm{T}} \quad \theta_1 \cdots \theta_{\mathrm{I}} \quad \omega]$.

According to the above definition, matrix A consists of four matrices where $A_1$is a $(|I| * |T|) \times (|I| * |T|)$ identity matrix. A_2 is a $|T| \times (|I| * |T|)$ matrix consisting of $|I|$ exclusive identity sub-matrices. A_3 is a $|I| \times (|I| * |T|)$ matrix consisting of $|I|$ exclusive

sub-matrices where each *i*th sub-matrix is composed by an all-ones vector as its *i*th row and zeros otherwise. A_4 is a unit row matrix of $|I| * |T|$ columns consisting of all 1s.

For further clarification, an example of $A\varepsilon \leq b$ for $|I| = 2$ and $|T| = 3$ follows. In this example, all sub-matrices are illustrated:

$$\begin{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \\ \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \end{bmatrix} \times \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix} \leq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \theta_1 \\ \theta_2 \\ \omega \end{bmatrix} \tag{57}$$

We can prove that the above matrix of coefficients (A) is a Totally Unimodular matrix. Therefore, we can claim that all optimal ε_{ij}s in our model are binary variables. A Totally Unimodular matrix is a matrix such that every square sub-matrix has a determinant of 1,0 or -1. According to Hoffman and Kruskal (Hoffman & Kruskal, 1956), if A is an integral matrix, then A is totally unimodular if and only if for every integral vector b, the vertices of the polyhedron {x : x ≥ 0 and Ax ≤ b} is integral. Matrix $\begin{bmatrix} A_2 \\ A_3 \end{bmatrix}$ is a Totally Unimodular (TU) matrix according to Hoffman and Kruskal (1965) since it contains only 0 and 1 elements, each column contains at most two nonzero elements and it can be partitioned (as A_2 and A_3) in such a way that nonzero elements with the same sign fall into the separate partitions. Since $\begin{bmatrix} A_2 \\ A_3 \\ 0 \end{bmatrix}$ is totally unimodular, we can perform a pivot operation by multiplying each row of A_3 by 1 and then add to the last row. By repeatedly doing this for each row of A_3, we obtain the matrix $\begin{bmatrix} A_2 \\ A_3 \\ A_4 \end{bmatrix}$. According

to the properties of TU matrices by which the matrix obtained by pivoting any row of a TU matrix is TU, we can conclude that the resulting matrix $\begin{bmatrix} A_2 \\ A_3 \\ A_4 \end{bmatrix}$ is still TU. Similarly following the properties of TU matrices, according to which, the matrix obtained by appending the identity matrix to a TU is TU, since A_1 is an identity matrix, therefore $A = \begin{bmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \end{bmatrix}$ is also a TU matrix.

Since A is proved to be a TU matrix, we can conclude that the optimal ε_{ij} can take only integer values and therefore they are 0 or 1 in our settings.

Discussion and Conclusion

In this research, we provided an in-depth look into the conditions under which the QCH surgical unit operates. We analyzed the data for two fiscal years 2012 and 2013 in order to better understand the dynamics among elective patient arrivals, required procedures, surgeons` specification with regards to patients length of stay, elective and emergency patients` occupancy level in the ward and the effect of performing surgeries by different surgeons on different days of the week on the ward occupancy level.

To address the first research question and determine which master surgery schedule provides the minimum peak occupancy in the ward, we initially developed a MIP model in order to minimize the peak load by re-arranging the master surgery schedule for surgeons over a 28-day horizon. We were able to provide the hospital with an optimal master schedule that reduces maximum bed occupancy in the ward by 19% and reduces the occupancy variation by 57% while maintaining the same OR availability as under the current practice of the hospital.

Furthermore, with regards to the second research question, we considered the length of stay of patients as the main source of fluctuation in ward occupancy. We demonstrated the significant reduction in peak load (compared to current practice) in the ward by optimizing the block schedule by incorporating the stochasticity of LOS. We tried one scenario of our model using the same LOS distribution for all surgeons. The distribution was derived by aggregating the data of LOS of all patients for all surgeons. Thus, we demonstrated that failing to consider variations in LOS between surgeons lead to an average maximum occupancy in the ward of 31.79 beds, while by incorporating different LOS distribution for each surgeon the maximum drops to 27.09 beds. Other parameters of

the model, although they could be considered as a fluctuation source, such as the number of inpatient referrals per block for each surgeon, were considered as constant at their average level.

We also expanded our model in such a way as to maintain a consistent block schedule for each surgeon per surgeons` preference to respond to our third research question. We demonstrated that accommodating this preference did not greatly diminish the advantage of the optimized schedule over current practice.

Different scenarios were developed to evaluate the performance measure for the ward under varying OR capacity levels. Our main hypothesis is verified as new schedules were demonstrated to significantly reduce the variation in bed occupancy in the ward over the course of the block schedule. Other scenarios including ORs during weekends suggest further smoothing in the ward occupancy is possible by providing different master schedules.

Our findings show that additional OR availability above the minimum requirement does not much affect the peak load occupancy in the ward but does result in greater fluctuations in OR capacity requirements from one day to the next. Moreover, opening ORs during the weekends smooths out the OR utilization level throughout the scheduling horizon; as such opening 3 ORs per day per weekend reduces the variation in OR utilization by 62.5%.

We applied simulation in order to verify our MIP models and also in order to have a broader overview of the system including other units and other patient streams. By simulation, we included the interaction among elective and emergency surgical patients and medical patients. Moreover, simulation provides a longer horizon perspective compared to the MIP and can accommodate fluctuations over months and thus better

reflect reality. The simulation results also suggest lower maximum occupancy and less variation at ward by new schedules.

Additionally, and to respond to the last two research questions, we developed different RO models based on different scenarios for the allowable deviation from the average inpatient referral rate to the ward in order to accommodate uncertainty in the number of inpatients referred per block per surgeon as the other source of fluctuation in the ward occupancy level. The main objective in developing RO models is to guard against potential periods of high congestion. The robust schedules were shown to provide better performance (lower peak load in the ward) under periods of congestion without too much deterioration in the performance under periods of average load. Thus, according to our findings, it is worthwhile for the hospital to develop the master surgery schedule and estimate their bed requirements in the ward using the robust models in order to better prepare for periods of congestion.

Direction of future research

Hospitals are in need of more data and information driven operational plans rather than relying on intuition. Therefore, our attempt could be considered as an initial step to further improvements at the hospital.

The current achievements can be expanded to maintain one important property in the surgical schedule to share the weekends among different surgeons. Currently, the hard constraints force the schedule to be repetitive as much as possible. Therefore, the model is incentivized to use the same surgeons on the weekend each week. However, this won`t necessarily be acceptable to those surgeons that are forced to work on the weekends. We need to apply a strategy in our modeling to rotate the weekend allocation among different surgeons while maintaining the improved peak load in the ward.

Although we don`t observe any meaningful seasonality in the data, it is favorable for the hospital to have a longer horizon in the picture as a cycle. This maintains the same concern of Blake J. T. et al. (2002) that although the master surgical schedule must be cyclic it ought also to be flexible.

With regards to Operating Rooms, we can include OR specifications for special surgeries into the modeling process. Moreover, optimizing OR utilization can be another objective to the model to better address utilization issues in the surgical unit.

The robust model can accommodate more flexibility and level of uncertainty by redefining the uncertainty set not only for each surgeon but per surgeon per day of surgery. By this approach, we can impose other levels of control such as allowable deviation per day, limit on the number of blocks in which a surgeon deviates during a cycle and/or the total possible surgeon-day deviation.

After the master surgical schedule, the next important concern of hospital management would be the scheduling of elective patients into the block. This concern addresses different aspects involved in the problem such as patients waiting time and urgency of surgery, surgeons and resources idle time, as well as overtime. In further studies, adding the within block scheduling to the model could be of value to the hospital.

References

Addis, B., Carello, G., & Tànfani, E. (2014). A Robust Optimization Approach for the Operating Room Planning Problem with Uncertain Surgery Duration. *Proceedings of the International Conference on Health Care Systems Engineering*, 175-189.

Ben-Tal, A., El Ghaoui, L., & Nemirovski, A. (2009). *Robust Optimization.* Prinston University Press.

Rachuba, S., & Werners, B. (2014). A robust approach for scheduling in hospitals using multiple objectives. *Journal of the Operational Research Society*, 546–556.

Astaraky, D., & Patrick, J. (2015). A Simulation based Approximation Dynamic Programming Approach to Multi-class, Multi resource Surgical Scheduling. *European Journal of Operational Research*, 309-319.

Beliën , J., & Demeulemeester , E. (2007). Building cyclic master surgery schedules with leveled resulting bed occupancy. *European Journal of Operational Research*, 1185–1204.

Beliën. J, & Demeulemeester, E. . (2009). A decision support system for cyclic master surgery scheduling with multiple objectives. *Journal of Scheduling*, 147–161.

Bertsimas, D., & Sim, ·. (2003). Robust discrete optimization and network flows. *Math. Program*, 49-71.

Blake JT, Dexter F, & Donald J . (2001). Operating room manager' s use of integer programming for assigning block time to surgical groups: a case study. *Anesthesia & Analgesia*, 143–148.

Bortfeld, T., Chan, T.C.Y., Trofimov, A., & Tsitsikl, J.N. (2008). Robust management of motion uncertainty in intensitymodulated radiation therapy. *Operations Research*, 1461–1473.

Carnes, T, & Price D. . (2011). An optimization framework for smoothing surgical bed census via strategic block scheduling. *Manufacturing Service Opeation Management*, 488–494.

Chow, V., Puterman, M., Salehirad, N., Huang, W., & Atkins, D. (2011). Reducing surgical ward congestion through improved surgical scheduling and uncapacitated simulation. *Production and Operations Management*, 418–430.

D. C. Lane, C. M. (2000). Looking in the Wrong Place for Healthcare Improvements: A System Dynamics Study of an Accident and Emergency Department. *Operational Research Society*, 518-531.

D. C. Lane, C. M. (2000). Looking in the Wrong Place for Healthcare Improvements: A System Dynamics Study of an Accident and Emergency Department. *Operational Research Society*, 518-531.

D. C. Lane, C. Monefeldt, & J. V. Rosenhead. (1998). Looking in the wrong place for health care improvements. *The Journal of the Operational Research Society*, 518-531.

D.N., P., & A, K. (2008). Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, 1011–1025.

David C. Lane, D., Monefeldt, C., & Rosenhead, J. (2000). ooking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Operational Research Society*, 518-531.

Denton, B. T., Miller, A. J., Balasubramanian, H. J., & Huschka, T. R. (2010). Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty. *Operations Research*, 802-816.

Gabrel, V., Murat , C., & Thiele , A. (2014). Recent Advances in Robust Optimization: An Overview. *European Journal of Operational Research*, 471–483.

Gupta, D. (2007). Surgical suites'operations management. *Production and Operations Management*, 689–700.

Hoffman, A., & Kruskal, J. (1956). Integral Boundary Points of Convex Polyhedra. *Linear Inequalities and Related Systems, Annals of Mathematics Studies*, 223–246.

Holte M., & Mannino C. . (2012). The implementer/adversary algorithm for the cyclic and robust scheduling problem in health-care. *European Journal of Operational Research*, 551-559.

Jun, J. B, Jacobson, S. H, & Swisher, J. R. (1999). Application of discrete-event simulation in health care clinics: A survey. *Journal of the operational research society*, 109-123.

Lane, D., Monefeldt, C., & Husemann, E. (2003). Lane, D. C., C. Monefeldt, et al. Client involvement in simulation model building: hints and insights from a case study in a London hospital. *Health care management science*, 105-116.

Lane, D., Monefeldt, C., & Husemann, H. (2003). Client involvement in simulation model building: hints and insights from a case study in a London hospital. *Health Care Management Science*, 105–116.

McCrory, B., LaGrange, C., & Hallbeck, M. (2014). Quality and Safety of Minimally Invasive Surgery: Past, Present, and Future. *Biomedical Engineering and Computational Biology*, 1-11.

Min, D. , & Yih, Y. (2010). Scheduling elective surgery under uncertainty and downstream capacity constraints. *European Journal of Operational Research*, 642–652.

Olivares M, , Terwiesch C, & Cassorla L. (2008). Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Science*, 41–55.

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

Santibanez, P., M. Begen, & D. Atkins . (2007). Surgical block scheduling in a system of hospitals: An application to resource and wait list management in a British Columbia health authority. *Health Care Management Science*, 269-282.

Shylo O.V, Prokopyev O. A, & Schaefer A. J. (2013). Stochastic Operating Room Scheduling for High-Volume Specialties Under Block Booking. *INFORMS Journal on Computing*, 682-692.

Timothy C.Y. Chan , & Velibor V. Mišić . (2013). Adaptive and robust radiation therapy optimization for lung cancer. *European Journal of Operational Research*, 745–756.

Van Oostrum J.M.x, Van Houdenhoven M., & Hurink J.L. (2008). A master surgery scheduling approach for cyclic scheduling in operating room departments. *OR Spectrum*, 355–370.