University of California

Los Angeles

Resource Planning Models for Healthcare Organizations

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Management

by

Sandeep Rath

2016

Abstract of the Dissertation

Resource Planning Models for Healthcare Organizations

by

Sandeep Rath
Doctor of Philosophy in Management
University of California, Los Angeles, 2016
Professor Kumar Rajaram, Chair

In this dissertation I look at two problems of resource planning at two major healthcare organizations. The Greater Los Angeles Station of the Veterans Health Administration and the UCLA Ronald Reagan Medical Center.

The first chapter of this thesis is a brief introduction to the research presented in subsequent chapters. The second chapter of this thesis considers the problem of minimizing daily expected resource usage and overtime costs across multiple parallel resources such as anesthesiologists and operating rooms, which are used to conduct a variety of surgical procedures at large multi-specialty hospitals. To address this problem, a two-stage mixed-integer stochastic dynamic programming model with recourse is developed. The first stage allocates these resources across multiple surgeries with uncertain durations and prescribes the sequence of surgeries to these resources. The second stage determines actual start times to surgeries based on realized durations of preceding surgeries and assigns overtime to resources to ensure all surgeries are completed using the allocation and sequence determined in the first stage. A data driven robust optimization method that solves large-scale real-sized versions of this model close to optimality is developed. This model is validated and implemented as a decision support system at the UCLA Ronald Reagan Medical Center. This has led to an average daily cost savings of around 7% or estimated to be $2.2 million on an annual basis.

In addition, the insights based on this model have significantly influenced decision making at the operating services department at this hospital.

In the third chapter of this thesis the planning problem for HIV screening, testing, and care is analyzed. This problem consists of determining the optimal fraction of patients to be screened in every period as well as the optimum staffing level at each part of the healthcare system to maximize the total health benefits to the patients measured by quality-adjusted life-years (QALYs) gained. This problem is modeled as a nonlinear mixed integer programming program comprising disease progression (the transition of the patients across health states), system dynamics (the flow of patients in different health states across various parts of the healthcare delivery system), and budgetary and capacity constraints. On applying the model to the Greater Los Angeles (GLA) station in the Veterans Health Administration system, it was found that a Centers for Disease Control and Prevention recommended routine screening policy in which all patients visiting the system are screened for HIV irrespective of risk factors may not be feasible because of budgetary constraints. Consequently, the model was used to develop and evaluate managerially relevant policies within existent capacity and budgetary constraints to improve upon the current risk based screening policy of screening only high risk patients. Our computational analysis showed that the GLA station can achieve substantial increase (20% to 300%) in the QALYs gained by using these policies over risk based screening.

The fourth chapter of this thesis concludes with some remarks on future research.

The dissertation of Sandeep Rath is approved.

Felipe Caro

Uday S. Karmarkar

Aman Mahajan

Kumar Rajaram, Committee Chair

University of California, Los Angeles

2016

Table of Contents

List of Figures

List of Tables

Acknowledgments

My dissertation would not have been possible without the guidance, help and support from my committee members. I am extremely grateful to my advisor Prof. Kumar Rajaram who has been instrumental in my development as a scholar. Through our collaboration, and my experience as a Teaching Assistant for his courses, I have learned to express my ideas more clearly. I believe this well help me in the future to be a good researcher and an effective teacher. Prof. Aman Mahajan has provided me with great support and advice during my work related to UCLA Ronald Reagan Medical Center. Prof. Uday Karmarkar and Prof. Felipe Caro have been very helpful with their suggestions and advice during the course of my dissertation.

I also wish to thank Profs. Chris Tang, Guillaume Roels, Reza Ahmadi, Sarang Deo, Elisa Long and Fernanda Bravo for their mentorship and support during my PhD. A word of thanks to Prof. Ram Bala and Prof. Aditya Jain for providing me a starting point in my academic career at ISB. I would also like to acknowledge the support provided by Lydia Heyman and Craig Jessen from the UCLA Anderson PhD office. Also, I grateful to Dr. Michael Sopher, Dr. Ira Hofer and Dr. Catherine Duda for help at multiple stages of the UCLA Ronald Reagan Project.

My stay in Los Angeles would have been much less interesting without the presence of my friends. I would particularly like to thank my fellow PhD students at UCLA; Christian Blanco, Nimesh Patel, Aparupa Das Gupta, Priya Mittal and Paul Rebeiz. Also my long standing friends, Nitin Basant, Joy Deep Nath, Rashmi Vadnagare, Rizi Mohanty, Lavanya Manohar and Anita Venkiteswaran have provided me advice and help at times I have needed it most along with that much needed perspective of the world outside of academia.

Finally, I wish to convey my deepest gratitude to my mother. Her belief in my abilities and unfailing support has always encouraged me to do more with my

life.

Vita

2006	B.Tech. (Electrical Engineering), M.Tech. (Instrumentation Engineering) Indian Institute of Technology, Kharagpur.
2006–2007	Assistant Operations Manager, ITC Ltd. Chennai, India
2007–2010	Manufacturing Engineer, Schlumberger Ltd., Pune, India
2010–2011	Research Associate, Indian School of Business, Hyderabad, India
2011–2015	Teaching Assistant, UCLA Anderson School of Management, Los Angeles, California

Publications

Deo, S., Rajaram, K., Rath, S., Karmarkar, U. S., & Goetz, M. B. (2015). Planning for HIV Screening, Testing, and Care at the Veterans Health Administration. *Operations Research*, 63(2), 287-304

CHAPTER 1

Introduction

In healthcare organizations there is a growing demand for the use of operations management to create integrated systems that better utilize resources for improving outcomes while reducing costs (Green, 2012). The rapid adoption of modern data storage technologies by healthcare organizations (Bresnick, 2015) has created opportunities to improve operational effectiveness and derive insight by bringing together previously disconnected sources of data, namely, clinical, operational, pharmaceutical and patient behavior data (Groves et al., 2013). The development of such integrated solutions will require the use of novel predictive and prescriptive models. My research is a step in this direction.

In my research I look at resource planning problems at two large healthcare organizations. The Greater Los Angeles Station of the Veterans Health Administration and the UCLA Ronald Reagan Medical Center. I conclude with remarks on future research directions.

The second chapter of this thesis looks at assignment and scheduling for specialized resources for surgical suites at UCLA Ronald Reagan Medical Center. UCLA Ronald Reagan Medical Center (RRMC) is large multi specialty hospital conducting more than 26,000 elective surgeries annually across 12 specialties. UCLA RRMC had recently invested in a clinical record system and was keen on using the data generated to improve operational decision making. After discussions with the Operating Services Department at UCLA RRMC, we formulated a research project to develop a model-based decision support system which would aid the resource planners in their day to day operations.

On a daily basis the planners at UCLA RRMC assign operating rooms, anesthesiologists and nursing staff to surgeries. They also provide a schedule for performing the surgeries. The objective is to minimize the daily resource usage and staffing cost. This assignment and scheduling task is complex due the high degree of uncertainty in surgical procedures, the simultaneous use of multiple specialized resources, and the large scale of operations at the UCLA RRMC. To address this problem, we develop a two-stage stochastic dynamic programming model with recourse. The first stage allocates these resources across multiple surgeries with uncertain durations and prescribes the sequence of surgeries to these resources. The second stage determines actual start times to surgeries based on realized durations of preceding surgeries and assigns overtime to resources to ensure all surgeries are completed using the allocation and sequence determined in the first stage. We develop a data driven robust optimization method that solves large-scale real-sized versions of this model close to optimality. To improve the quality of solution, we apply a model calibration procedure based on historical data of surgery durations and surgical characteristics. We validate and implement this model as a decision support system at the UCLA Ronald Reagan Medical Center. This has led to an estimated average daily cost savings of around 7% or $2.2 million on an annual basis.

In addition, this model provided managerial insights to the operating services department. We find that small reduction in variance on surgical durations leads to significant reduction in total resource utilization costs while larger reductions in variability only leads to marginal levels of cost reduction. This suggests that rather than investing in costly capital intensive equipment to radically reduce variability in surgical durations, low-cost incremental efforts such as checklists and standardized protocols derived from best practices would be more beneficial. Such measures have also been advocated by surgeons (Gawande, 2010). We also find, similar to literature related to value to flexibility in manufacturing (Jordan and Graves, 1995), making only a few additional rooms multi-specialty reduces the cost significantly.

The third chapter of this thesis looks at planning for screening, treatment and care at the Greater Los Angeles Station of the Veterans Health Administration. In 2006 the Centers for Disease Control (CDC) had recommended a policy of routine screening for HIV, in which all patients arriving at a healthcare system would be offered an HIV detection test. This was done to aid in the early detection of HIV. The Greater Los Angeles (GLA) Station of Veterans Health Administration (VHA) wanted to investigate the feasibility of routine screening under current budgetary and resource constraints and develop alternate policies if necessary. In this chapter we developed a large-scale mixed-integer non-linear programming model incorporating disease progression, patient flow, service level constraints and budgetary constraints. The routine screening policy was found to be infeasible under the current budget level. Consequently, we developed alternate policies for screening and staffing policies using this model in order to maximize the Quality Adjusted Life Years (QALYs) gained across the GLA station at the current budget level. Using the model we also derived an estimate of the additional budget required for implementation of routine screening.

In addition to providing recommended policies for staffing and screening, this research also showed that though a policy such as routine screening may be cost effective from a aggregate population perspective, it may be infeasible while implementing in an organization due to budgetary constraints. Our analysis demonstrated that joint optimization of screening and staffing leads to better coordination between these two tasks. This leads to a higher proportion of infected patients getting connected to treatment earlier, resulting in a significant improvement of QALYs. The model insights have significantly influenced decision making process at this station of the VHA.

CHAPTER 2

Integrated Anesthesiologist and Room Scheduling for Surgeries

2.1 Introduction

Surgical procedures are complex tasks requiring the use of several specialized and expensive resources. In a hospital, the operating services department is responsible for managing resources used in surgical procedures. Every day, this department assigns to each surgery an operating room, an anesthesiologist, a nursing team and the requisite surgical materials. The department also determines the sequence in which these surgeries will be performed and the scheduled start times. While performing these actions, the department ensures that the cost of the operating room suite is minimized by reducing resource usage and overtime costs.

The operating services departments at large hospitals devote significant amount of time in making these resource management decisions. The complexity of these decisions is to due the following four primary reasons. First, operating room resources are expensive (Macario, 2010) and in short supply (Orkin et al., 2013) and thus surgeries are performed in highly resource constrained environments. Second, surgical procedures are often very specialized. Therefore, equipment and facility requirements govern whether a procedure can be performed in a particular room. Anesthesiologist assignments too are dictated by specialty. Studies have demonstrated that not only do surgeons often prefer to have an anesthesiologist of required sub-specialty (Ghaly, 2014), outcome indicators of surgical procedures are significantly better when anesthesia is delivered by an anesthesiologist with experience in that particular sub-specialty (McNicol, 1997). Pardo (2014) predict that increasingly anesthesiologists will be assigned by their sub-specialties. Third, the durations of surgical procedures are very difficult to predict (Kayis et al., 2012). This is because there are many procedures and newer procedures are constantly being developed (AMA, 2013). Consequently, historical data on all these procedures is not available. Furthermore, surgeon's estimates of durations are often unreliable. Studies have demonstrated systematic underestimation as well as overestimation of procedure times by surgeons while scheduling surgeries. Some surgeons overestimate the duration when they do not have enough cases to fill their scheduled block time while others may underestimate the time when they wish to fill in more cases (Laskin et al., 2013). Finally, the scale of large hospitals, in terms of the number of operating rooms, procedures conducted, the number and types of equipment and anesthesiologists used, makes the simultaneous scheduling of multiple resources a computationally challenging task.

I was exposed to these complexities in our work at the operating services department of the UCLA Ronald Reagan Medical Center (RRMC), a large multi-specialty hospital which consistently ranks amongst the best five hospitals in the United States (Roxanne Moster, 2014). The management of this department felt that the daily resource allocation decision played a significant role in overall department cost and in the service quality delivered to patients. They believed that these aspects can be significantly improved by developing an analytical model based approach that considered the key complexities in this environment and applied historical surgical data to decide resource assignment and scheduling. This paper describes the development, implementation and evaluation of a model based decision support system that uses a data driven robust optimization procedure to determine the daily scheduling of rooms and anesthesiologists for elective surgeries at UCLA RRMC.

There is a large body of literature on elective surgery scheduling. Min and Yih (2010) consider scheduling elective surgeries under uncertainty in surgery durations and downstream capacity constraints. Gupta (2007) discusses the broader issues of managing operating room suites in the context of elective surgeries. Reviews of literature related to operating room scheduling can be found in Blake and Carter (1996), Cayirli and Veral (2003) and Guerriero and Guido (2011). In their review Cardoen et al. (2010a) categorize operating room scheduling literature by patient characteristics, performance criteria, decision level, type of analysis, solution technique and whether the papers incorporate uncertainty or not. Below, we primarily focus on research that accounts for uncertainty in surgery duration, as required in our application.

The literature for scheduling operating rooms under uncertain surgery durations has primarily been focused on single resource type corresponding to the operating room. Green and Savin (2008) model a single operating room appointment scheduling model with no shows using a queuing approach with Poisson arrivals and deterministic service times. Denton and Gupta (2003) develop a stochastic optimization model to optimize start times for procedures with random durations while Mancilla and Storer (2012) develop heuristics to find near optimal sequencing of surgeries in a single operating room. Mak et al. (2014a) solve the appointment scheduling and sequencing in a single operating room following a robust optimization procedure. They find that the widely used heuristic of ordering the jobs by variance is optimal under mild conditions. Denton et al. (2010) solve the problem of assignment of surgeries to multiple parallel operating rooms under fixed costs of operating rooms and variable overtime costs. However, none of these papers consider multiple resources, the simultaneous sequencing and start times of surgeries or are tested with data in a large-scale application context. While a single resource type may be sufficient for specialized surgery suites, multi-specialty hospitals like the UCLA RRMC require a holistic solution of surgery scheduling that simultaneously optimizes on all specialized parallel and multiple resources.

Literature related to multiple resource types is relatively scarce. Beliën and Demeulemeester (2008) and Meskens et al. (2013) consider integrated operating room scheduling with multiple resources under deterministic surgery durations. Batun et al. (2011) consider scheduling of surgeries given two resource types: operating rooms and surgeons under stochastic surgery durations. However, they do not consider specializations of rooms and anesthesiologists, and consider a problem significantly smaller in scale than in our application. For the scale of problem at UCLA RRMC, sample average approximation based stochastic optimization procedures as used in Denton and Gupta (2003) and Min and Yih (2010) were intractable. This is due to the large number of possible integer assignments in the first stage which increases the number of samples required to achieve convergence in objective value and solution. These difficulties in employing this method has been described in a more general context by Kleywegt et al. (2002). Furthermore, the overall complexity of the large-scale problem in our application precluded finding even feasible solutions using leading commercially available solvers such as `ddsip` (Märkert and Gollmer, 2008) that employ the state-of-the-art procedures for solving stochastic programs such as dual decomposition methods (Carøe and Schultz, 1999). In order to circumvent these problems A robust optimization procedure (Bertsimas et al., 2013; Bertsimas and Thiele, 2006) is used. While a similar approach has been used by Denton et al. (2010) and Mak et al. (2014b), our work extends theirs by considering multiple resource types. This extension requires significant modification to existing solution methods.

Our paper makes the following contributions. First, we consider two types of parallel resources which are of critical importance to specialties: operating rooms and anesthesiologists and we simultaneously optimize their assignment and sequencing. Second, we develop an efficient solution method using robust optimization to provide effective solutions to large-scale problems. An important element when applying robust optimization is the estimation of an uncertainty set. We develop an estimation procedure to estimate the sets using historical data. This data driven robust optimization approach was successful in solving the full scale problem for the entire surgery suite at the RRMC within 10 minutes with a performance gap within 5% from the lower bound. Third, our methodology significantly outperforms the best benchmark procedures in the literature. Fourth, we develop a model based decision support system which has been validated and implemented at the UCLA RRMC. This has considerably improved upon current practice and has resulted in average daily cost savings of around 7% or estimated to be $2.2 million on an annual basis. Further, the insights from our work has had a notable impact on decision making at the hospital.

The remainder of the paper is organized as follows. §2.2 provides a detailed problem description at the UCLA RRMC. §2.3 presents the model formulation, properties and solution procedure. In §2.4 we describe the procedure for parameter estimation and model calibration. §2.5 provides the results of the computational analysis. §2.6 describes the implementation of this model at the UCLA RRMC, presents the financial benefits, provides managerial insights and describes the organizational impact of this work.

2.2 Problem Description

Operating services is one of the largest departments at the UCLA RRMC with around $120 million in annual revenues representing about 10% of this hospital's revenues. This department serves around 27,000 patients annually by conducting around 2700 *types of* elective and emergency surgical procedures across 12 specialties. Emergency surgeries are conducted in 3 dedicated operating rooms with a separate team of anesthesiologists. Since emergency surgeries are separated from elective surgeries and account for only about 15% of revenues, the management of this department asked us to focus solely on elective surgeries. To perform these surgeries, the operating services department uses 23 Operating Rooms (ORs), which are further divided across these 12 specialties that require specific equipment. General surgery procedures can be performed in any of these 23 ORs dedicated for the exclusive use of elective surgeries. The details on the number of rooms that can perform each specialty is provided in Table 2.1. Surgeries are scheduled to start in operating rooms only between 7 am and 3 pm. Further, there are fixed costs for opening an operating room each day. This consists of an initial cleaning and equipment setup costs along with daily nurse and technician staffing costs, whose assignments do not depend on specialty. In addition, overtime costs are incurred for nurses and technicians if the rooms are required to be open beyond 3 pm. Finally, these operating rooms are scheduled and staffed simultaneously.

Table 2.1: Summary of resource by specialty

Surgery Specialty	Number of ORs Available	Number of Anesthesiologists Available
Vascular	1	9
Neuro	3	10
Plastics	23	NA
ENT	23	NA
Urology	23	NA
Liver	1	8
Thoracic	2	5
Cardiac	3	14
Trauma	1	NA
Pediatric	2	12
Eye Surgery	23	NA
General	23	NA

During our work here, there were 92 anesthesiologists at the UCLA RRMC divided across these 12 specialties. The number of anesthesiologists by specialty is also shown in Table 2.1. Anesthesia for surgeries in some specialties can be administered by any anesthesiologist. Such specialties are denoted by NA in this table. There are three shifts of equal duration for the anesthesiologists: day (7 a.m till 3 p.m.), late (11 a.m. till 7 p.m), and night (7 p.m till 3 a.m). Each anesthesiologist is preassigned to exactly one shift and thus, the regular working hours for each anesthesiologist is eight hours. The assignment of anesthesiologists to surgeries is according to the specialty and availability. In addition, anesthesiologists can only be assigned to surgeries that begin during their shift. Overtime costs for anesthesiologists are incurred if surgeries in progress exceed the duration of the shift. A certain number of anesthesiologists who are not scheduled to work on a given day are asked to be on standby or on call, so that they can be called to work if necessary. However, when anesthesiologists are assigned from call, there are significant costs for using such an option. Anesthesiologists assigned from call do not incur overtime costs. The anesthesiologists on call are informed of their status the previous day and assigned surgeries the following day as required.

It is important to note that in the context of large multi-specialty hospitals such as the UCLA RRMC, surgeons are not part of the operating services department. They are usually from the independently administered specialty departments at this hospital and on some occasions can be from other hospitals. The surgeons bring their patients and use the operating services department as a service provider. Thus, the operating services department does not have the option of assigning surgeons to patients. For this reason, we assume that each surgery-surgeon combination is already set and we consider them together. This ensures that each surgery has a clear and unchangeable link to the surgeon. This aspect is also consistent with the literature in this area (Dexter and Traub, 2002; Marques et al., 2014).

Typically, a request to schedule a surgery is initiated by the surgeon on behalf of the patient with general admissions at the hospital. This request is assigned a date based on the earliest availability in the block reservations for the particular specialty. Once all the elective surgery requests have been received the day before the surgery, the operating services department decides which operating room to open, finalizes assignment of these rooms and anesthesiologist to surgeries, determines start times of surgeries and effectively specifies the sequence of all the surgeries. Add-on surgeries are not considered here as depending on availability, they are assigned to the operating rooms dedicated to emergency procedures. Define utilization as the fraction of the available shift time that is used by a particular resource. Inefficient assignment and scheduling of anesthesiologists and operating rooms to surgeries leads to low utilization and overtime of these resources. As seen in Figure 2.1, average daily utilization across the anesthesiologists is close to 0.75, with around 25% of days having an average daily utilization of less than 0.70. However, despite these lower levels of utilization, the average number anesthesiologists on call is around 6 per day. Similarly, for operating rooms the average daily utilization is close to 78% (Figure 2.2) but the average overtime per day is around 18 hours. Taken together average call and overtime costs for anesthesiologists and rooms at this department are about 33% of revenues. A more effective assignment and scheduling system could potentially reduce overtime and on call costs.

Figure 2.1: Histogram of average daily utilization of anesthesiologists

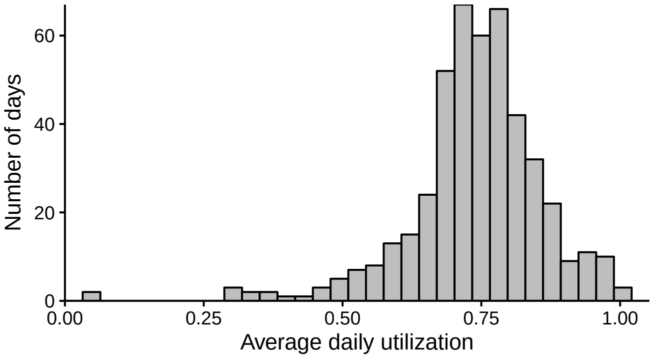
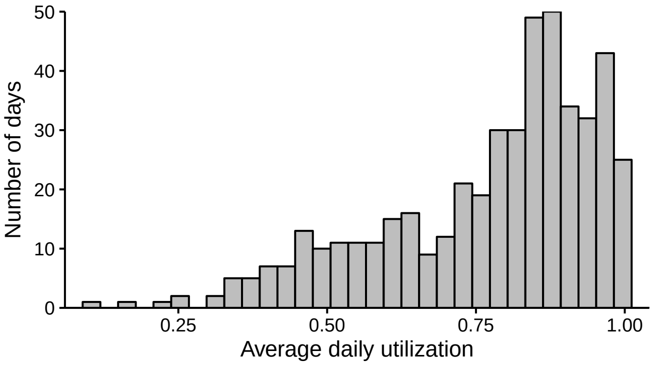
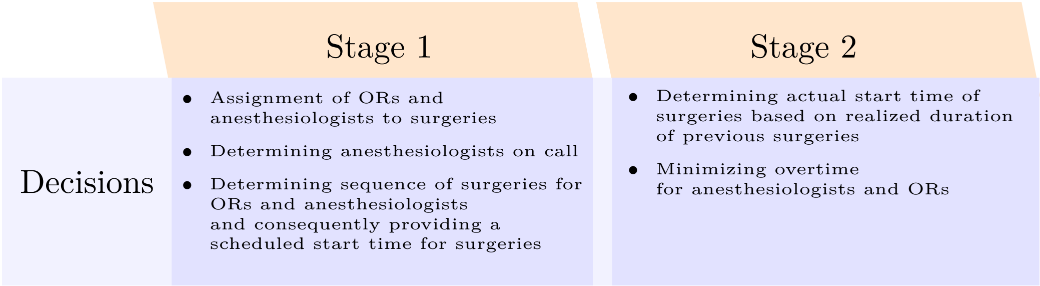


Figure 2.2: Histogram of average daily utilization of operating rooms



Assignment and scheduling decisions at this hospital are complicated by the large number of ORs and anesthesiologists, variety in surgical procedures, variability in anesthesiologist workload, and unpredictability in surgery durations. More details on these aspects are provided in the Appendix. These complicating aspects make any long term resource planning inadequate on the *day* of the surgery. Thus, the management of the operating services department felt that the daily expected resource usage and overtime costs across operating rooms and anesthesiologists could be considerably lowered by developing an optimization model, which led to our involvement. This model is formulated as a two-stage mixed-integer stochastic dynamic program with recourse. The first stage of this model allocates these resources across multiple surgeries with uncertain durations and prescribes the sequence of surgeries to these resources. Assuming that each surgery should be scheduled as early as possible, this consequently provides a scheduled start time for surgeries. The second stage determines the actual start times to surgeries based on realized durations of preceding surgeries and assigns overtime to resources to ensure all surgeries are completed using the allocation and sequence determined in the first stage. The stages of the model are shown in Figure 2.3. The size and complexity of the problem precluded solution using conventional methods. Therefore, we develop a data driven robust optimization approach that solves large-scale real-sized versions of this model close to optimality. Next, we describe the model formulation, present its properties and describe its solution techniques.

Figure 2.3: Model stages and decisions



	Stage 1	Stage 2
Decisions	• Assignment of ORs and anesthesiologists to surgeries • Determining anesthesiologists on call • Determining sequence of surgeries for ORs and anesthesiologists and consequently providing a scheduled start time for surgeries	• Determining actual start time of surgeries based on realized duration of previous surgeries • Minimizing overtime for anesthesiologists and ORs

2.3 Model

We start by presenting a model formulation of the integrated anesthesiologist and room scheduling problem for surgeries. To provide a precise definition of the model, let $h, i, j \in I$ index the set of surgeries, $a \in A$ index the set of anesthesiologists, and $r \in R$ index the set of operating rooms. We define the following variables.

x_{ia}: 1 if anesthesiologist a is assigned to surgery i, 0 otherwise

y_a: 1 if anesthesiologist a is assigned from call, 0 otherwise

z_{ir}: 1 if room r is assigned to surgery i, 0 otherwise

v_r: 1 if room r is assigned any surgery, 0 otherwise

u_{ij}: 1 if surgery i precedes surgery j, 0 otherwise

α_{ija}: 1 if surgery i and j are assigned to anesthesiologist a and i precedes j, 0 otherwise

β_{ijr}: 1 if surgery i and j are assigned to room r and i precedes j, 0 otherwise

s_i: Scheduled start time of surgery i (hrs)

S_i: Actual start time of surgery i (hrs)

$Over_a$: Overtime of anesthesiologist a (hrs)

$Over_r$: Overtime of room r (hrs)

In addition, let $\mathbf{x} = (x_{ia})$ $\forall i \in I, a \in A$, $\mathbf{y} = (y_a)$ $\forall a \in A$, $\mathbf{z} = (z_{ir})$ $\forall i \in I, r \in R$, $\mathbf{u} = (u_{ij})$ $\forall i, j \in I$, $\mathbf{s} = (s_i)$ $\forall i \in I$ denote the vectors associated with these variables. Next we define the following parameters:

κ_{ia}^{A}: 1 if anesthesiologist a can be assigned to surgery i, 0, otherwise

κ_{ir}^{R}: 1 if surgery i can be done in room r, 0 otherwise

w_a: 1 if anesthesiologist a is on call, 0 otherwise

c_r: Fixed cost of opening operating room r (\$/day)

c_{oa}: Overtime cost of anesthesiologist a (\$/hr)

c_{or}: Overtime cost of room r (\$/hr)

c_q: Cost of assigning anesthesiologist from call (\$/day)

t_a^{start}: Start time of shift associated with anesthesiologist a (hr)

t_a^{end}: End time of shift associated with anesthesiologist a (hr)

T^{end}: End time of the day (hr)

$M,\ M_{seq}, M_{anesth}, M_{room}$: large positive numbers

The durations d_i of surgery i is uncertain $\forall i \in I$ and can be considered as a random variable. The vector of surgery durations for the day is represented by $\mathbf{d} = (d_i),\ \forall i \in I$. We incorporate the uncertainty in surgery durations through a robust optimization approach where we model d_i as an uncertain parameter that takes values in $[\bar{d}_i - \hat{d}_i, \bar{d}_i + \hat{d}_i]$, where $\bar{d}_i$ is the nominal duration for surgery i and $\hat{d}_i$ is one sided maximum deviation for surgery i. We define the scaled deviations of d_i about its nominal value as $f_i = (d_i - \bar{d}_i)/\hat{d}_i$. Note that the scaled deviation f_i can take a value in $[-1, 1]$. Following the approach in Bertsimas and Sim (2004) and Denton et al. (2010) we subject the scaled deviations to a constraint $\sum_{i \in I} |f_i| \leq \tau$ so that, the total deviation across all surgeries is less than a known threshold τ. Here, τ bounds the total maximum deviation of surgery duration from the nominal value across all surgeries. This threshold is called the budget of uncertainty and represents the level of pessimism on the number of surgeries deviating from their nominal value. If $\tau = 0$, it is equivalent to solving the nominal value problem with $d_i = \bar{d}_i,\ \forall i \in I$.

The Integrated Anesthesiologist and Room Scheduling Problem (IARSP) consists of two stages. The first stage problem assigns rooms and anesthesiologists to surgeries, and prescribes a sequence of surgeries to be performed in each room and by each anesthesiologist. The second stage recourse function determines actual start times to surgeries based on realized durations of preceding surgeries and assigns overtimes to resources such that all the surgeries are completed in the assignment and sequence prescribed by the first stage problem. Here, the two-stage approach assumes that all information about actual surgery durations is known early in the morning, which is of course not the case. However, this simplification has no impact on the solution since the only recourse action is to accumulate overtime without changing the sequences. Further, this simplification is consistent with the literature on surgery scheduling employing two-stage stochastic models with recourse (Denton et al., 2010; Batun et al., 2011; Mancilla and Storer, 2012). The *IARSP* can be written as:

$$(IARSP) \qquad \mathcal{V}^*(\tau) = \min \left\{ \sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a + \boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}) \right\} \tag{2.1}$$

subject to,

$$\sum_{a \in A} x_{ia} = 1 \qquad \forall i \in I \tag{2.2}$$

$$\sum_{r \in R} z_{ir} = 1 \qquad \forall i \in I \tag{2.3}$$

$$z_{ir} \leq v_r \qquad \forall i \in I, r \in R \tag{2.4}$$

$$x_{ia} \leq v_a + y_a \qquad \forall i \in I, a \in A \tag{2.5}$$

$$y_a \leq w_a \qquad \forall a \in A \tag{2.6}$$

$$s_i \geq t_a^{start} - M(1 - x_{ia}) \qquad \forall i \in I, a \in A \tag{2.7}$$

$$x_{ia} \leq \kappa_{ia}^A \qquad \forall i \in I, a \in A \tag{2.8}$$

$$z_{ir} \leq \kappa_{ir}^R \qquad \forall i \in I, r \in R \tag{2.9}$$

$$\alpha_{ija} \leq u_{ij} \qquad \forall i, j \in I, a \in A \tag{2.10}$$

$$\beta_{ijr} \leq u_{ij} \qquad \forall i, j \in I, r \in R \tag{2.11}$$

$$u_{ij} + u_{ji} \leq 1 \qquad \forall i, j \in I \tag{2.12}$$

$$\alpha_{ija} + \alpha_{jia} \geq x_{ia} + x_{ja} - 1 \qquad \forall i, j \in I, a \in A \tag{2.13}$$

$$\beta_{ijr} + \beta_{jir} \leq z_{ir} \qquad \forall i, j \in I, r \in R \tag{2.14}$$

$$\beta_{ijr} + \beta_{jir} \leq z_{jr} \qquad \forall i, j \in I, r \in R \tag{2.15}$$

$$\beta_{ijr} + \beta_{jir} \geq z_{ir} + z_{jr} - 1 \qquad \forall i, j \in I, r \in R \quad (2.16)$$

$$\alpha_{ija} \geq x_{ia} + x_{ja} + \beta_{ijr} - 2 \quad \forall i, j \in I, r \in R, a \in A \quad (2.17)$$

$$\beta_{ijr} \geq z_{ir} + z_{jr} + \alpha_{ija} - 2 \quad \forall i, j \in I, r \in R, a \in A \quad (2.18)$$

$$x_{ia}, y_a, z_{ir}, , u_{ij}, v_r, \alpha_{ija}, \beta_{ijr} \in \{0, 1\} \qquad \forall i, j \in I, r \in R, a \in A \quad (2.19)$$

$$s_i \geq 0 \qquad \forall i \in I \quad (2.20)$$

Objective function (2.1) consists of three terms. The first term is the fixed cost for opening operating rooms each day. The second term is the cost of assigning anesthesiologists from call. The third term $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$, represents the worst case second stage cost and is described in detail below. Constraints (2.2) and (2.3) assign each surgery exactly one anesthesiologist and one operating room respectively. Constraint (2.4) ensures that v_r is set to 1 whenever any surgery is assigned to operating room r. Constraint (2.5) ensures that an anesthesiologist can be assigned to a surgery only if they are on regular duty or on call. Constraint (2.6) enforces that an anesthesiologist can be assigned from call only if they are listed in the call list. Constraint (2.7) ensures that an anesthesiologist can be assigned a surgery only if the scheduled start time of the surgery is after the shift start time of the anesthesiologist. Constraints (2.8) and (2.9) ensure that surgeries are assigned rooms and anesthesiologists by specialty. Constraint (2.10) enforces the condition that if an anesthesiologist is used to conduct surgery i before surgery j, then surgery i has to precede surgery j or u_{ij} is set to 1. Constraint (2.11) imposes the similar condition and sets u_{ij} to 1 when surgery i precedes surgery j in an operating room Constraint (2.12) ensures that only one of u_{ij} or u_{ji} can be 1. Constraint (**??**) is required to maintain consistency of schedule between any three surgeries that follow each other, so that if i precedes j and j precedes h, then i should precede h. Constraints (**??**)-(2.13) restrict that only one of α_{ija} and α_{jia} is 1 only if both surgeries i and j are assigned to anesthesiologist a. This also ensures that the sequencing constraints for anesthesiologists α_{ija} is active only for those surgeries that are assigned to the same anesthesiologist. Constrains (2.14)-(2.16) are similar logical constraints corresponding to the sequencing of rooms. Constraints (2.17) and (2.18) maintain consistency of sequencing variables between operating rooms and anesthesiologists. Constraint (2.17) enforces that if anesthesiologist a and operating room r is assigned to surgeries i and j and i precedes j in operating room r then i has to precede j in assignment to anesthesiologist a. Constraint (2.18) is a similar constraint that makes sure that if surgery i precedes surgery j with anesthesiologist a, then i has to precede j in the assignment of operating room r. Constraints (2.19) and (2.20) represent variable domains.

The worst case second stage cost is given by:

$$\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max_{\mathbf{d}\in\mathcal{D}(\tau)} \boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}) \tag{2.21}$$

$$\boldsymbol{\mathcal{D}}(\tau) = \left\{\mathbf{d}\in\mathbb{R}^{|I|} : d_i = \bar{d}_i + f_i\hat{d}_i, i\in I, \mathbf{f}\in\boldsymbol{\mathcal{F}}(\tau)\right\} \tag{2.22}$$

$$\boldsymbol{\mathcal{F}}(\tau) = \left\{\mathbf{f}\in\mathbb{R}^{|I|} : \sum_{i\in I}|f_i| \le \tau, -1 \le f_i \le 1\right\} \tag{2.23}$$

$\boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$ is the total overtime cost across all resources, for a given assignment, sequence, scheduled surgery start times and surgery durations. This is maximized over the vector of surgery durations $\mathbf{d}$ to determine the worst case cost, where $\mathbf{d}$ is restricted to lie in the uncertainty set $\mathcal{D}(\tau)$ given by (2.22). This equation restricts d_i, the duration of surgery i, to lie within a maximum deviation of $\hat{d}_i$ from the nominal value of the duration $\bar{d}_i$. The total extent of such deviations is specified by the set $\mathcal{F}(\tau)$, which is defined by (2.23) and is well suited to our problem context. In particular, the effective allocation of multiple parallel resources such as anesthesiologists and rooms which are used repeatedly across the surgeries in a given day requires a specification of τ, an overall level or budget of uncertainty across surgical durations. This is enforced by (2.23), which specifies that the maximum deviation across all surgeries is at most τ. A schedule based on a large τ would be overly accommodating towards the second stage cost, while a schedule corresponding to a small τ would not be accommodating enough.

In §2.4, we present a methodology to determine $\bar{d_i}$, $\hat{d_i}$ and τ based on historical data.

In determining $\boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d})$, it is important to note that the only decision variables at this stage are the actual start times of the surgeries and the overtime for the anesthesiologists and rooms. We pick these variables to minimize total overtime costs while ensuring that all the surgeries scheduled for the day are completed and there are no conflict in actual start times of surgeries assigned to the same resource. To compute $\boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d})$ we formulate the linear program:

$$\boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d}) = \min \left\{ \sum_{a \in A} c_{oa} Over_a + \sum_{r \in R} c_{or} Over_r \right\} \tag{2.24}$$

subject to,

$$S_j \geq S_i + d_i - M_{seq}(1 - u_{ij}) \qquad \forall i, j \in I \tag{2.25}$$

$$S_i \geq s_i \qquad \forall i \in I \tag{2.26}$$

$$Over_a \geq S_i + d_i - t_a^{end} - M_{anesth}(1 - x_{ia} + y_a) \qquad \forall i \in I, a \in A \tag{2.27}$$

$$Over_r \geq S_i + d_i - T^{end} - M_{room}(1 - z_{ir}) \qquad \forall i \in I, r \in R \tag{2.28}$$

$$S_i, Over_a, Over_r \geq 0 \qquad \forall i \in I, a \in A, r \in R \tag{2.29}$$

The objective function consists of the sum of overtime across all the resources. Constraint (2.25) ensures that the start time of the succeeding surgery is only after the end time of the preceding surgery. Constraint (2.26) ensures that the actual start time of the surgery can be no earlier than the scheduled start time. Constraints (2.27) and (2.28) define the overtime for anesthesiologists on regular duty and operating rooms respectively, which is the time difference between the end time of the last surgery in that shift and the regular shift end time for the resource. Constraint (2.29) restricts start time and overtime variables to be non-negative variables. The *IARSP* is a robust optimization model with the recourse represented by this linear program. We next develop some structural properties that is useful in constructing solution techniques for this model.

Proposition 1. *The* IARSP *has relatively complete recourse.*

All proofs are provided in the appendix. Proposition 1 implies that for every feasible first stage solution there exists a feasible second stage solution. This proposition allows us to evaluate second stage costs for every feasible first stage solution. This is important for the solution method for the *IARSP* described in §2.3.1.1. However, evaluating $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ for any given first stage solution requires one to solve the problem given in equations (2.21)-(2.29), which is not easy due to the max-min operator in its objective. The following proposition simplifies the computation of $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ and consequently the *IARSP*.

Proposition 2. *If parameter τ is chosen to be a positive integer, then $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ can be reformulated as the following mixed integer program.*

$$
\begin{aligned}
\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}) = \max \Bigg\{ & \sum_{i \in I} \left(\bar{d}_i \pi_i + \xi_i \hat{d}_i \right) + \sum_{i \in I} s_i \phi_i - M_{seq} \sum_{i,j \in I i \neq j} \lambda_{ij} \left(1 - u_{ij}\right) \\
& - M_{anesth} \sum_{\substack{i \in I \\ a \in A}} \mu_{ia} \left(1 - x_{ia} + y_a\right) \\
& - M_{room} \sum_{\substack{i \in I \\ r \in R}} \theta_{ir} \left(1 - z_{ir}\right) - \sum_{\substack{i \in I \\ r \in R}} \theta_{ir} T^{end} - \sum_{\substack{i \in I \\ a \in A}} \mu_{ia} t_a^{end} \Bigg\}
\end{aligned}
$$

subject to,

$$\sum_{i \in I} \mu_{ia} \leq c_{oa} \qquad \forall a \in A \tag{2.30}$$

$$\sum_{i \in I} \theta_{ir} \leq c_{or} \qquad \forall r \in R \tag{2.31}$$

$$\sum_{\substack{j \in I \\ j \neq i}} \lambda_{ij} - \sum_{\substack{j \in I \\ j \neq i}} \lambda_{ji} + \sum_{a \in A} \mu_{ia} + \sum_{r \in R} \theta_{ir} - \phi_i \geq 0 \qquad \forall i \in I \tag{2.32}$$

$$\sum_{j \in I - \{i\}} \lambda_{ij} + \sum_{a \in A} \mu_{ia} + \sum_{r \in R} \theta_{ir} = \pi_i \qquad \forall i \in I \tag{2.33}$$

$$\sum_{i \in I} f_i \leq \tau \tag{2.34}$$

$$\xi_i \leq M_f f_i \qquad \forall i \in I \tag{2.35}$$

$$\xi_i \leq \pi_i \qquad \forall i \in I \tag{2.36}$$

$$\xi_i, \pi_i, \lambda_{ij}, \theta_{ir}, \mu_{ia}, \phi_i \geq 0 \qquad \forall i, j \in I, a \in A, r \in R \tag{2.37}$$

$$f_i \in \{0, 1\} \qquad \forall i \in I \tag{2.38}$$

As described in the appendix, the proof of this proposition follows from strong duality of the second stage recourse problem for a given first stage solution. Note that as a consequence of Proposition 2 in which τ is set to an integer, $f_i, \forall i \in I$ are also now binary variables. Thus, in the worst case, surgeries are set to either their nominal value or maximum positive deviation. In effect, τ can now be interpreted as the upper bound on the number of surgeries that reach their maximum deviation. Thus, restricting τ to be a positive integer as in this proposition allows for a more natural interpretation of τ, which was important in the application context. As discussed in §2.4, this interpretation drives our data driven method in setting a parametric value for τ from historical data. Propositions 1 and 2 imply that $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ can be evaluated for any given first stage feasible solution by solving a mixed integer program. In particular, let $(\boldsymbol{\pi}^l, \boldsymbol{\xi}^l, \boldsymbol{\lambda}^l, \boldsymbol{\mu}^l, \boldsymbol{\theta}^l, \boldsymbol{\phi}^l)$ be the solution to $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ for some given $(\mathbf{x}^l, \mathbf{y}^l, \mathbf{z}^l, \mathbf{u}^l, \mathbf{s}^l)$. Then, the following propositions provide a lower bound and characterize the structure of $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$. They will be used in the solution method provided in §2.3.1.1.

Proposition 3. *A lower bound on* $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ *is provided by* $\sum_{i \in I} \left(\bar{d}_i \pi_i^l + \xi_i^l \hat{d}_i \right) + \sum_{i \in I} s_i \phi_i^l - M_{seq} \sum_{i,j \in I i \neq j} \lambda_{ij}^l \left(1 - u_{ij}\right) - M_{anesth} \sum_{i \in I a \in A} \mu_{ia}^l \left(1 - x_{ia} + y_a\right) - M_{room} \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l \left(1 - z_{ir}\right) - \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l T^{end} - \sum_{\substack{i \in I \\ a \in A}} \mu_{ia}^l t_a^{end}$.

Proposition 4. $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ *is a piecewise-linear convex function in the first stage decision variables* $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}$*.*

In light of Proposition 4, the *IARSP* now reduces to a piecewise-linear convex mixed integer program in which $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$, the convex part of the objective function can be evaluated by using Proposition 2 and solving a mixed integer program. However, given this non-linearity and the large number of integer variables in our application, the *IARSP* cannot be solved using powerful solvers for non-linear programs such as BARON (Sahinidis, 2014) and DICOPT (Viswanathan and Grossmann, 1990). Consequently, we develop the following model based heuristic procedure to solve this problem.

2.3.1 Solution Methods

We start by describing the model based heuristic. We then present the process currently being employed at the hospital (i.e., the practitioner's heuristic). Finally, we discuss sample average approximation based techniques that are commonly used in literature, which we use to benchmark the model based and practitioner's heuristics. The performance of these methods will be discussed in §2.5.

2.3.1.1 Model Based Heuristic.

This heuristic is based upon Kelley's algorithm (Kelley, 1960) as described in Thiele et al. (2009) to solve robust optimization problems with recourse. Here, we consider the *IARSP* and in light of Proposition 4, approximate $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ by a piecewise linear equation via successive linear cuts. We then use this approximation in constructing the master problem at the k^{th} iteration of the heuristic, $MP(k)$, defined as:

$MP(k)$

$$\min_{v_r, y_a, \phi} \left\{ \sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a + \psi \right\} \tag{2.39}$$

subject to,

$$(2.2) - (2.20)$$

$$\psi \geq \sum_{i \in I} \left(\bar{d}_i \pi_i^l + \xi_i \hat{d}_i \right) + \sum_{i \in I} s_i \phi_i^l - M_{seq} \sum_{i,j \in I i \neq j} \lambda_{ij}^l \left(1 - u_{ij}\right)$$

$$- M_{anesth} \sum_{i \in I a \in A} \mu_{ia}^l \left(1 - x_{ia} + y_a\right) \tag{2.40}$$

$$- M_{room} \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l \left(1 - z_{ir}\right) - \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l T^{end} - \sum_{\substack{i \in I \\ a \in A}} \mu_{ia}^l t_a^{end} \qquad l = 0, 1, 2..., k-1 \tag{2.41}$$

$$\psi \geq 0 \tag{2.42}$$

Observe that in this problem, we approximate the value of $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ by a variable $\psi \geq 0$. To improve this approximation, in each iteration of the heuristic we use (2.41) to enforce the condition that ψ is greater than or equal to the lower bound of $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$, as established by Proposition 3. This results in constraints (2.41) in which ψ approximates $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ by a piecewise linear equation via successive linear cuts. An integral part of the heuristic is determining a good solution to the $MP(k)$ at each iteration of the computation, as the quality of this first stage solution will affect the convergence time for this heuristic. The following proposition is useful in developing a good quality first stage solution.

Proposition 5. *If two surgeries i, j are assigned the same operating room and the same anesthesiologist and $\bar{d}_i \leq \bar{d}_j$, $\hat{d}_i \leq \hat{d}_j$, then $u_{ij}^* = 1$, where u_{ij}^* is the optimal value of u_{ij} in the optimal solution of* IARSP.

This proposition implies that surgeries with smaller nominal values and smaller maximum deviations are performed before other surgeries when they share anesthesiologists and operating rooms. This is intuitive as scheduling the longer and more uncertain surgery upfront would have disruptive effect on the subsequent surgeries that share resource with this surgery. This would lead to a higher overtime cost. Therefore, to preclude this in any optimal solution, $u^*_{ij} = 1$. This proposition will be used in the model based heuristic formalized by the following algorithm.

Model Based Heuristic

Step 1. Initialize $U \leftarrow \infty, L \leftarrow 0, k \leftarrow 0, l \leftarrow 0$. Set $\epsilon > 0$ to be sufficiently small.

Step 2. Solve the *MP(k)* and let the solution be $\tilde{x}^k_{ia}, \tilde{y}^k_a, \tilde{z}^k_{ir}, \tilde{u}^k_{ij}, \tilde{s}^k_i, \tilde{v}^k_r, \tilde{\alpha}^k_{ija}, \tilde{\beta}^k_{ijr}, \tilde{\psi}^k$. **If** $i, j \in I$ satisfy the conditions of Proposition 5, **then** set $\tilde{u}^k_{ij} \leftarrow 1, \tilde{u}^k_{ji} \leftarrow 0$. Set $L \leftarrow \sum_{r\in R} c_r \tilde{v}^k_r + \sum_{a\in A} c_q \tilde{y}^k_a + \tilde{\psi}^k$.

Step 3. Compute $\boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}}^k, \tilde{\mathbf{y}}^k, \tilde{\mathbf{z}}^k, \tilde{\mathbf{u}}^k, \tilde{\mathbf{s}}^k)$ for given $\tilde{x}^k_{ia}, \tilde{v}^k_r, \tilde{z}^k_{ir}, \tilde{u}^k_{ij}, \tilde{s}^k_i$, obtained in **Step 2** by solving the mixed integer programming formulation given in Proposition 2. Let the optimal solution be $\xi^k_i, \pi^k_i, \lambda^k_{ij}, \theta^k_{ir}, \mu^k_{ia}, \phi^k_i$. $U \leftarrow \min\left\{U, \sum_{r\in R} c_r \tilde{v}^k_r + \sum_{a\in A} c_q \tilde{y}^k_a + \boldsymbol{\mathcal{Q}}(\tilde{\mathbf{x}}^k, \tilde{\mathbf{y}}^k, \tilde{\mathbf{z}}^k, \tilde{\mathbf{u}}^k, \tilde{\mathbf{s}}^k)\right\}$.

Step 4. If $U - L < \epsilon$ go to **Step 6**, **else** go to **Step 5**.

Step 5. $k \leftarrow k + 1$. Add constraint

$$\psi \geq \sum_{i\in I}\left(\bar{d}_i \pi^k_i + \xi_i \hat{d}_i\right) + \sum_{i\in I} s_i \phi^k_i - M_{seq} \sum_{i,j\in I i\neq j} \lambda^k_{ij}\left(1 - u_{ij}\right)$$

$$-M_{anesth} \textstyle\sum_{i\in I a\in A} \mu^k_{ia}\left(1 - x_{ia} + y_a\right)$$

$$-M_{room} \textstyle\sum_{\substack{i\in I\\ r\in R}} \theta^k_{ir}\left(1 - z_{ir}\right) - \textstyle\sum_{\substack{i\in I\\ r\in R}} \theta^k_{ir} T^{end}$$

$$- \textstyle\sum_{\substack{i\in I\\ a\in A}} \mu^k_{ia} t^{end}_a$$

to the *MP(k)*. Go to **Step 2**.

Step 6. $\tilde{x}^k_{ia}, \tilde{y}^k_a, \tilde{z}^k_{ir}, \tilde{u}^k_{ij}, \tilde{v}^k_r, \tilde{s}^k_i, \tilde{\alpha}^k_{ija}, \tilde{\beta}^k_{ijr}$ is the heuristic solution to *IARSP*.

The above algorithm is well suited for the IARSP as from Proposition 4, its objective is piecewise-linear convex and cutting plane methods such as those used in Step 5 of this algorithm are finitely convergent for piecewise linear functions

(Ruszczyński, 2006). Also, in this algorithm, in early iterations, the solution in Step 2 is obtained by employing the *user callbacks* feature of the solver used to solve $MP(k)$. Here, instead of solving this problem to optimality in the initial iterations, we request the solver to return a feasible solution, which is then used to apply cuts in Step 4 and approximate the convex function $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ at each corresponding feasible solution. We do so because *MP(k)* is a problem with many integer variables and solving it to optimality can be computationally expensive with poor returns at early iterations when $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ has not been approximated well enough by constraints (2.41). A potential drawback with this approach is that the initial feasible solutions obtained might be poor and many cuts might be required before the stopping criteria is reached. However, Proposition 5 is used in Step 2 to improve the quality of the feasible solution and ensure faster convergence. Results on the computational performance and the time required for this heuristic is provided in Section §2.5.

2.3.1.2 Practitioner Heuristic.

This heuristic is based on the current planning process used at the operating service department of the UCLA RRMC. Such types of heuristics have been reported in the literature (Dexter and Traub, 2002; Cardoen et al., 2010b). The practitioner's heuristic consists of the following steps.

Step 1: Assign surgeries to operating rooms in sequential fashion in order of start times requested by the surgeons, by surgery specialty and duration estimates from surgeons, until the last surgery in the room can start before the end of the shift for the operating room.

Step 2: Assign one anesthesiologist to each room such that the anesthesiologist can perform most of the surgeries in the room.

Step 3: A few anesthesiologists are assigned to surgeries across rooms in order to ensure all surgeries have been assigned an anesthesiologist by specialty.

Step 4: If above plan cannot be implemented by anesthesiologists on regular duty, assign anesthesiologists from call.

2.3.1.3 Benchmark Heuristic.

Here we use Sample Average Approximation (SAA) based methods similar to those provided in Denton et al. (2007) to benchmark the model based heuristic. In SAA, instead of using the worst case formulation we solve with expected second stage costs. The expectation is based on scenarios drawn from an estimated distribution of surgeries. The resultant two-stage stochastic optimization problem is solved by the *L-Shaped* method (Birge and Louveaux, 1988). We describe the model formulation of the SAA version and provide details on the estimation of the distribution in the appendix. However, since the sample average based method was unable to solve large-scale problems such as those found in the application, we use this method on smaller problems constructed from real data. We provide the results of the comparison between SAA based method and the robust optimization based method in §2.5.

2.4 Parameter Estimation and Model Calibration

In this section we use historical data of surgery durations to choose the uncertainty sets $\boldsymbol{\mathcal{D}}(\tau)$ and $\boldsymbol{\mathcal{F}}(\tau)$. The performance of robust optimization depends closely on the definition of these uncertainty sets. If the optimal values of $d_i,\ \forall i \in I$ in the inner maximization problem ($\max_{\mathbf{d}\in\mathcal{D}(\tau)} \boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$) are significantly larger than the corresponding $\bar{d}_i$, the resulting first stage problem will be overly pessimistic towards the realization of surgery durations and lead to higher first stage costs. Conversely, if the optimal values of $d_i,\ \forall i \in I$ are too close to the corresponding $\bar{d}_i$, the uncertainty sets would not cover many cases of future realizations in which the surgery durations deviate significantly from the nominal value and this would lead to higher second stage costs. Thus, we need to look at the combined first and second stage costs while designing the uncertainty sets. Designing the uncertainty sets involves setting the following parameters: the nominal surgery duration $\bar{d}_i,\ \forall i \in I$, the maximum deviation $\hat{d}_i,\ \forall i \in I$, and the robust optimization parameter τ.

There have been several approaches suggested for designing uncertainty sets. Ben-Tal et al. (2009) provide the theoretical background for deciding good uncertainty sets. Denton et al. (2010) use the 10th and 90th percentile width of historical surgery durations as the width $[\bar{d}_i - \hat{d}_i, \bar{d}_i + \hat{d}_i]$ in an operating room assignment application. Subsequently, they perform sensitivity analysis and calibrate their model to an equivalent SAA based solution to decide the robust optimization parameter τ. Bertsimas et al. (2013) propose using statistical hypothesis tests to construct uncertainty sets. Denton et al. (2010) and Bertsimas et al. (2013) model the uncertainty sets based on historically observed values of a *single* uncertain parameter. In our application with a wide variety of surgery specialties with considerable variability in surgery durations across specialties, a percentile width not conditional on surgery characteristics would be unnecessarily wide, leading to an overly pessimistic uncertainty set. Therefore, we incorporate these characteristics and propose a joint estimation and calibration procedure to design the uncertainty set. Our procedure provides tight uncertainty sets that take into account observable surgery characteristics while making no assumptions on the probability distribution of surgery durations. We further calibrate the uncertainty set by evaluating the performance of the robust solution to empirical realizations.

There were two data sets available to us. The first data set $\Delta^E = \left\{\widetilde{d}^{(m)}, \widetilde{\mathbf{b}}^{(m)}\right\}_{m=1}^{M}$ consists of $M = 25700$ samples of $\widetilde{d}^{(m)}$ corresponding to the historical realization of durations of surgery m and $\widetilde{\mathbf{b}}^{(m)}$ which represents the observed characteristics of surgery m. Table 2.2 provides details on the surgery characteristics included in $\widetilde{\mathbf{b}}(m)$, $\forall m$ and the variable names used for the subsequent regression. The second data set Δ^C was partitioned into disjunctive training and testing sets, $\Delta^{C-Train} = \left\{\widetilde{\mathbf{d}}^{(n)}, \widetilde{\mathbf{b}}^{(n)}\right\}_{n=1}^{N_1}$ and $\Delta^{C-Testing} = \left\{\widetilde{\mathbf{d}}^{(n)}, \widetilde{\mathbf{b}}^{(n)}\right\}_{n=1}^{N_2}$. For these data sets, $N_1 = 120$ days and $N_2 = 60$ days. Both these data sets consist of $\widetilde{\mathbf{d}}^{(n)}$ representing the vector of realized durations and $\widetilde{\mathbf{b}}^{(n)}$ denoting the vector of surgery characteristics of all surgeries performed on day n.

Table 2.2: Surgery data Description

Surgery characteristics	Description	Variable Name
Realized surgery duration	In hours	ACTUALHRS
Surgeon's estimate of surgery duration	In hours	BOOKEDHRS
Patient Class	Inpatient, Outpatient or Same Day Admit	PATCLASS

Continued on next page

Table 2.2 – continued from previous page

Surgery characteristics	Description	Variable Name
Booked Current Procedural Terminology (CPT) code	Medical code maintained by American Medical Association defines the services to be performed during surgery. A surgery may have multiple CPT codes. The surgeon provides a list of services that maybe performed as a part of the surgery. The realized CPT codes may and often do vary from the booked CPT code. Surgeries in our data set covered 2700 unique CPT codes.	CPT
ASA Score	A system for assessing fitness of patients before surgery, higher number signifies a less fit patient. Takes integer values between 1 and 6.	ASA
Patient age	In years	AGE
Surgery Service	Cardiac Surgery, Neuro Surgery, etc., full list as in Table 2.11	SERVICE

Continued on next page

Table 2.2 – continued from previous page

Surgery characteristics	Description	Variable Name
Surgeon's name	Names of 493 surgeons unique surgeons (providers) who have performed surgeries in the period over which data was available	PROV
Number of CPT codes	Number of CPT codes associated with procedure	NUMCPT

We use Δ^E and $\Delta^{C-Train}$ to estimate $\boldsymbol{\mathcal{D}}(\tau)$ and $\boldsymbol{\mathcal{F}}(\tau)$. Most of the research in estimating uncertainty sets is for single stage problems when feasibility is not guaranteed. Since we have two stages and second stage feasibility is guaranteed by Proposition 1, we develop the following procedure that comprises an estimation and a calibration step.

Step 1: Estimation

First, for a given parameter $\rho \in (0, 1)$, we define conditional quantile functions $g_L(\mathbf{b}; \rho)$ and $g_U(\mathbf{b}; \rho)$ such that:

$$\mathbf{P}\left[\widetilde{d} \le g_L(\widetilde{\mathbf{b}}; \rho)\right] = \frac{1-\rho}{2}, \text{ and } \mathbf{P}\left[\widetilde{d} \ge g_U(\widetilde{\mathbf{b}}; \rho)\right] = \frac{1-\rho}{2}$$

Thus, given observed surgery characteristics $\widetilde{\mathbf{b}}$ the future realization $\widetilde{d}$ will lie in the set $\left[g_L(\widetilde{\mathbf{b}}; \rho), g_U(\widetilde{\mathbf{b}}; \rho)\right]$ with probability ρ. The true quantile functions are not known to us, so we obtain estimates of the quantile functions $\hat{g}_L(\widetilde{\mathbf{b}}; \rho)$ & $\hat{g}_U(\widetilde{\mathbf{b}}; \rho)$ through a conditional quantile regression method (Koenker, 2005) applied on the data set Δ^E. The use of conditional quantile regression for estimating uncertainty sets has been recently proposed by Tulabandhula and Rudin (2014). Quantile regression estimates the quantiles of the response variable (i.e., the surgery durations), given certain values of the predictor variables. Quantile regression has several advantages over the commonly used ordinary least squares (OLS) regression. First, this approach suits our application better since our objective is to find upper and lower bounds on surgery durations such that future realization would lie within this bound with a given probability. Conditional quantiles provide these bounds without making any assumption on the probability distribution of surgery durations. Second, quantile regression is more robust to outliers and third, it does not assume the dispersion of the response variable to be independent of the predictor variables.

We perform quantile regression using the `quantreg` package available in `R` (Koenker, 2013). The response variable is the realized surgery duration. The possible set of predictors are the surgeon's estimate of surgery duration, the fitness level of the patient prior to the surgery measured by the American Society of Anesthesiologist (ASA) score[1], the age of the patient, whether the patient is an inpatient or outpatient, the specialty of the surgery, the surgeons name, the services provided indicated by the type of the Current Procedure Terminology (CPT)[2] codes used and the number of CPT codes used by the surgeons. For the CPT codes and surgeon's names we cluster the variables via a k-means clustering similar to He et al. (2012). This clustering was done in order to account for the large number of factors in these variables and to avoid over-specification of the model. The details of the clustering procedure followed is provided in the Appendix.

The selection of variables was done comparing the Akaike Information Crie-

[1] https://www.asahq.org/resources/clinical-information/asa-physical-status-classification-system

[2] http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt.page

tria (AIC) and the Mean Square Prediction Error (MSPE). The results of these tests are provided in the Appendix. We also checked for collinearity using Variance Inflation Factors (VIF) following the criteria discussed in (Hair et al., 2006, pp. 191-193). Highly collinear variables (i.e. with VIF $\geq$ 10) were removed. For example, we found that the specialty of surgery had a VIF of 40.1 as it was collinear with CPT codes, as these codes were specific to a specialty. On performing these tests we found that the ASA score, the surgeon's estimate of duration, patient class (Inpatient, Outpatient or Same Day Admit), clustered variables corresponding to surgeons and CPT codes were significant. The ASA score is a strong indicator of increasing complexity of the surgical procedure since it is an indicator of the level of fitness of the patient before coming into the surgery. A patient with an ASA score of 1, implying the patient is a healthy person would be expected to demonstrate fewer complications during surgery while a patient with an ASA score of 3 (with severe systemic disease) would be expected to have more complications during surgery. We found that the coefficient of ASA score was -0.023 at the 0.1 quantile and 0.122 at the 0.9 quantile. Thus, every increase in ASA score contributes to approximately 7.3 minutes ($\approx$ 0.122 hrs) of additional surgery time at the 0.9 quantile level, while the effect of increment in ASA score is negligible for very short surgeries. This is intuitive as the negative effects of patient fitness would be significant for longer surgeries and would not be as impactful for shorter surgeries. As expected, the surgeon's estimate of surgery duration would be strongly correlated with the actual duration and would explain variance not captured by other variables, since there are several factors that the surgeon is aware of that are not captured by other available data. However, as explained previously there is some error in surgeon's estimates as well. We found that surgeon's estimate was on an average 12 minutes higher than actual surgery durations. Also, the coefficient of surgeons' estimates in the quantile regression model was smaller for shorter surgeries than for longer surgeries. This is because surgeons tend to be more accurate in their estimates for longer surgeries than for shorter surgeries. One possible explanation is that it was observed that surgeons tend to round to the nearest quarter of an hour while providing their estimates. This leads to an error which is more pronounced for shorter surgeries than for longer surgeries. We also found that clusters of surgeons are significant because as described in the Appendix these in effect represented the experience level of surgeons. Finally, as anticipated, the type of surgery itself with its associated CPT code affected surgical durations. However, the number of CPT codes was not significant as they could be associated with relatively simpler sub-procedures common to all surgeries. Similarly, the age of the patient was not significant as it was captured in the surgeons estimate of duration.

Once we have obtained the estimated conditional quantile functions, for each surgery $i \in I$ we set, $\bar{d}_i + \hat{d}_i = \hat{g}_U(\widetilde{\mathbf{b}}_i; \rho)$, $\bar{d}_i - \hat{d}_i = \hat{g}_L(\widetilde{\mathbf{b}}_i; \rho)$. This gives,

$$\bar{d}_i = \frac{\hat{g}_U(\widetilde{\mathbf{b}}_i; \rho) + \hat{g}_L(\widetilde{\mathbf{b}}_i; \rho)}{2} \text{ and } \hat{d}_i = \frac{\hat{g}_U(\widetilde{\mathbf{b}}_i; \rho) - \hat{g}_L(\widetilde{\mathbf{b}}_i; \rho)}{2}$$

Define $\tau' \in [0, 1]$ so that $\tau = \lfloor \tau' |I| \rfloor$. Here, τ' represents the fraction of total surgeries in a given day $|I|$ which have reach their maximum duration. We then substitute the above equations in (2.22) and (2.23) for observed surgery characteristics vector $\mathbf{b}_i$ and *given* parameters $\rho \in [0, 1]$ and $\tau' \in (0, 1)$. Then the uncertainty sets are given by:

$$\begin{aligned} \boldsymbol{\mathcal{D}}(\tau) = \boldsymbol{\mathcal{D}}(\rho, \tau') = \Bigg\{ \mathbf{d} \in \mathbb{R}^{|I|} : d_i = & \frac{\hat{g}_U(\widetilde{\mathbf{b}}_i; \rho) + \hat{g}_L(\widetilde{\mathbf{b}}_i; \rho)}{2} \\ & + f_i \frac{\hat{g}_U(\widetilde{\mathbf{b}}_i; \rho) - \hat{g}_L(\widetilde{\mathbf{b}}_i; \rho)}{2}, i \in I, \mathbf{f} \in \boldsymbol{\mathcal{F}}(\tau') \Bigg\} \end{aligned} \tag{2.43}$$

$$\boldsymbol{\mathcal{F}}(\tau) = \boldsymbol{\mathcal{F}}(\tau') = \left\{ \mathbf{f} \in \mathbb{R}^{|I|} : \sum_{i \in I} |f_i| \leq \lfloor \tau' |I| \rfloor, -1 \leq f_i \leq 1 \right\} \tag{2.44}$$

Step 2: Calibration

If we had full information on the surgery durations (i.e., if they were observable *ex ante*), and a deterministic solution could be executed, the resulting cost obtained when there is full information would be a lower bound to any heuristic solution. In stochastic programming this is referred to as the *wait-and-see* solution. The full information cost on day n is given as:

$$\mathcal{W}^{FI(n)} = \min \left\{ \sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a + \sum_{a \in A} c_{oa} Over_a + \sum_{r \in R} c_{or} Over_r \right\} \tag{2.45}$$

subject to,

$$(2.2) - (2.20)$$

$$(2.25) - (2.29)$$

We solve $\mathcal{W}^{FI(n)}$ for each day in $\Delta^{C-Train}$ with $\mathbf{d} = \widetilde{\mathbf{d}}^{(n)}$.

The first stage variables for day n obtained by solving *IARSP* for day n using the model based heuristic are $(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)})$. The cost of the model based heuristic under a realized duration vector $\widetilde{\mathbf{d}}^{(n)}$ is defined as:

$$\begin{aligned}\mathcal{W}(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)}; \rho, \tau', \widetilde{\mathbf{d}}^{(n)}) = &\sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a \\ &+ \boldsymbol{\mathcal{R}}\Big(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)}, \widetilde{\mathbf{d}}^{(n)}\Big)\end{aligned}$$

This represents the cost that would be realized at the end of day n if the model based heuristic was implemented with uncertainty set $\boldsymbol{\mathcal{D}}(\rho, \tau')$. The average performance of the model based heuristic across N samples relative to the full information case is defined as:

$$\bar{\mathcal{W}}(\rho, \tau') = \frac{1}{N_1} \sum_{n=1}^{N_1} \frac{\Big[\mathcal{W}(\mathbf{x}^{*(n)}, \mathbf{y}^{*(n)}, \mathbf{z}^{*(n)}, \mathbf{u}^{*(n)}, \mathbf{s}^{*(n)}; \rho, \tau', \widetilde{\mathbf{d}}^{(n)}) - \mathcal{W}^{FI(n)}\Big]}{\mathcal{W}^{FI(n)}} \tag{2.46}$$

We calculate $\bar{\mathcal{W}}(\rho, \tau')$ for several values of $\rho \in (0, 1)$ and $\tau' \in [0, 1]$ and choose the pair that minimizes $\bar{\mathcal{W}}(\rho, \tau)$. This is summarized in Table 2.3.

Table 2.3: $\bar{\mathcal{W}}(\rho, \tau')$ across ρ and τ'

ρ \ τ	0.1	0.2	0.3	0.4
0.8	1.52	1.43	1.48	1.57
0.85	1.43	1.35	1.45	1.59
0.9	1.41	1.27	1.44	1.59
0.95	1.38	1.24	1.4	1.55
0.98	1.38	1.29	1.42	1.57

From Table 2.3 we can see $\rho = 0.95$ and $\tau' = 0.2$ is optimal. This implies that at a 95% confidence level, we can set 20% of all surgeries to its maximum durations on any given day when we define the uncertainty sets $\boldsymbol{\mathcal{D}}(\tau)$ and $\boldsymbol{\mathcal{F}}(\tau)$ and solve the *IARSP*. At this value the model based heuristic solution was 24% more than the full information solution.

We used these values of ρ and τ' to evaluate the performance of model based heuristic relative to full information case as defined in (2.46) for the testing data set $\Delta^{C-Testing}$. Here, we found that the model based heuristic was 28% more than the full information solution. Thus, the out of sample performance, using $\Delta^{C-Testing}$ was close to the in sample performance using $\Delta^{C-Train}$. This provides validation to use these values of ρ and τ' in the computational analysis described next.

2.5 Computational Analysis

In this section, we conduct a computational analysis to evaluate our approach. To perform this analysis, we used data provided by the UCLA RRMC on all surgeries conducted in their operating room suite over a 14 month period. This analysis was essential to provide confidence in our method. Our computational analysis is divided into two sections. In §2.5.1 we evaluate the performance of the heuristic procedures described in §2.3.1. In §2.5.2 we compare the performance of the model based heuristic with the actual resource assignment and scheduling decisions made at this hospital and estimate the cost savings.

2.5.1 Performance Evaluation

The size and scope of the scheduling activities during this time period demonstrated considerable variation as shown in Table 2.4. To ensure that our computational analysis captured this range of variation, we constructed five problems of varying sizes as shown in Table 2.5. For each of these five sets of problem instances we considered different values for c_q, c_{or} and c_{oa}. The actual value of c_q, c_{oa} and c_{or} at UCLA RRMC were \$1000 per day, \$150 per hour, and \$450 per hour respectively. In addition to these actual values we considered values where we scaled one of these costs by a factor of 2 or 1/2 while keeping the other two at the current value. This led to 7 possible combinations of costs for each of the five problem instances and a total of $7 \times 5 = 35$ possible problems. We tried to solve the IARSP for these data sets using leading commercial solver for stochastic programs such as `ddsip` (Märkert and Gollmer, 2008). However, other than the smaller problem instances A and B, these solvers could not even generate feasible solutions after more than 24 hours of computation, and the runs were aborted. This provides validation for developing the model based heuristic to solve the *IARSP*.

Table 2.4: Data sets for performance analysis

	Number of surgeries conducted per day	Number of anesthesiologists working per day	umber of ORs functioning per day
Minimum	30	28	4
Maximum	62	38	23
Average	42	32	22
95 percentile	53	36	23

Table 2.5: Problems used for performance analysis

Instance	No. of surgeries, $\|I\|$	No. of rooms, $\|R\|$	No. of anesthesiologists, $\|A\|$
A	10	3	5
B	15	5	8
C	25	7	10
D	40	10	25
E	65	23	40

The heuristic procedures were coded in Python programming language (van Rossum, 2001). The computational analysis were run on a workstation with 3.8 GHz AMD A10 processor, 8 GB of RAM, and Linux Mint as the operating system. For the MIP subroutine calls we used Gurobi 5.63 (Gurobi Optimization, 2015) called from Python via the Gurobi Python Interface. In all the computations using the model based heuristic, we set the gap $\epsilon = 5\%$. Thus, all the solutions of the model based heuristic were within a 5% gap from the lower bound and were solved within 10 minutes.

Tables 2.6 and 2.7 summarize the results obtained for the computational analysis. In Table 2.6 the performance of the model based heuristic and the practitioner's heuristic procedure is compared with the cost of the SAA based solution for small-scale problems. This table shows that these procedures are all very close to the SAA method and this does not change with changes in the cost parameters. In Table 2.7 we consider the more realistic medium and large-scale problems. Since SAA is unable to solve these problems, we provide the performance of the model based heuristic with respect to the practitioner's heuristic. From Table 2.7, we note that for these problems the model based heuristic provides significant cost reductions over the practitioner's heuristic. In particular, the percentage cost reduction for these problems ranged from 2.26% to 7.56% averaging around 4.95%.

Table 2.6: Performance Evaluation of Heuristic Procedures for small scale problems

Instance	c_q	c_{oa}	c_{or}	% change in cost of Model Based Heuristic from SAA solution	% change in cost of Practitioner's heuristic solution from SAA solution
A	1000	150	450	0	2.97
	1000	150	900	0	4.61
	1000	150	225	0	1.15
	1000	300	450	0	2.34
	1000	75	450	0	0.75
	2000	150	450	0	2.97
	500	150	450	0	2.97
B	1000	150	450	-1.15	3.64
	1000	150	900	-1.74	5.67
	1000	150	225	0	1.35
	1000	300	450	-0.74	2.76
	1000	75	450	0	0.73
	2000	150	450	-1.15	3.64
	500	150	450	-1.15	3.64

We can also observe from Table 2.7 that the gains of the model based heuristic over the practitioner's heuristic improves as the size of the problem increases. This is because for smaller sized problems there are limited options and it is more likely the practitioner's heuristic achieves a solution that is close to optimal. Fur-

Table 2.7: Performance Evaluation of Heuristic Procedures for medium and large-scale problems

Instance	c_q	c_{oa}	c_{or}	% change in cost of Model Based Heuristic from Practitioner's heuristic solution
C	1000	150	450	-4.55
	1000	150	900	-6.45
	1000	150	225	-2.56
	1000	300	450	-3.46
	1000	75	450	-2.26
	2000	150	450	-5.55
	500	150	450	-4.57
D	1000	150	450	-6.64
	1000	150	900	-4.45
	1000	150	225	-3.37
	1000	300	450	-5.52
	1000	75	450	-2.65
	2000	150	450	-7.55
	500	150	450	-4.75
E	1000	150	450	-7.03
	1000	150	900	-5.74
	1000	150	225	-4.47
	1000	300	450	-6.69
	1000	75	450	-3.45
	2000	150	450	-7.56
	500	150	450	-4.75

ther, since in small sized problems $|I|$ is low, the number of surgeries that reach its worst case duration (i.e. $\tau = \tau'|I|$) is also low. In these circumstances, the solution of the model based heuristic and the practitioners heuristic are similar and close to the nominal value solution, where surgical durations are set to its nominal duration estimates. However, as the problem size increases, the number of surgeries reaching its worst case duration increases. Under these circumstances, the practitioner's heuristic is outperformed by the model based heuristic, as the optimization inherent to the model based heuristic is more effective in utilizing resources that can be shared across multiple specialties and procedures. Finally, note that increasing the on call costs leads to the practitioner's heuristic doing much worse as this heuristic opts for increasing the number of on call anesthesiologists rather than trading off on call costs against overtime costs.

2.5.2 Model Validation

The objective of model validation was to demonstrate that the model based heuristic provides tangible cost savings over current practice. This was an essential step in convincing the management of the Operating Services department to implement our method. We performed model validation in two stages. In the first stage the cost savings were computed using historical data. In the second stage, we conducted live validation, where we compared in real time our decisions with those made at this hospital. Note that while conducting these validations, the heuristic had precisely the same information the planners at the UCLA RRMC had at the point of planning.

In the historical validation, we took 80 sample days, such that we covered the range of problem sizes encountered. These 80 samples were divided into 5 sets as described in Table 2.8. We next calculated the average costs obtained by the model based heuristic and the costs resulting from the actual assignment and sequencing that was done by the RRMC planners. This reduction in costs across the five problems are also reported in Table 2.8 and this shows that the benefits of using the model based heuristic was significant and was increasing in problem size.

Table 2.8: Results from historical validation

Surgeries per day	% of days	% reduction of cost of Model Based Heuristic from cost of actual plan	Cost savings ($)
<10	28	0	
10-30	4	2.4	
30-40	18	3.3	
40-50	41	7.2	
50-65	9	8.9	

Table 2.9: Results from live validation

Surgeries per day	% of days	% reduction of cost of Model Based Heuristic from cost of actual plan	Cost savings ($)
<10	30	0	
10-30	0	0	
30-40	11	2.1	
40-50	48	6.4	
50-65	11	9.1	

The real-time live validation was conducted over a 4 week period. The number of surgeries per day over this period was similar to the range of problem sizes observed historically as is shown in Table 2.8. The results for the live validation is given in Table 2.9. This table shows that our heuristic reduced costs from current practice on average from 6.4% to 9.1% on 16 out of the 28 days corresponding to weekdays which were not holidays. This implied an estimated annual cost savings between \$2 million and \$2.86 million. It is also important to note that the practitioners heuristic and the model based heuristic provided the same solution in the weekends where the number of surgeries conducted are low and both these methods provided solutions corresponding to the nominal value solution.

The model based heuristic outperforms the current practice in both historical and live validation for the following reasons. First, the nominal values obtained via quantile regression procedure provided a better predictors for the realized surgical duration than the surgeon's estimates. Second, on average, around 50% of the surgeries exceeded the nominal value. This required an increase in realized workload from the nominal workload. We found that this increase can be effectively achieved by setting $\tau' = 20\%$. This led to the model based heuristic operating with fewer operating rooms and fewer anesthesiologists than actually used at the hospital, since these resource assignments were based on trading off the fixed costs for these resources with the chance of incurring overtime. Third, on average in 60% of the surgeries, the surgeon's estimate of surgery duration exceeded the realized duration. The planners chose additional resources and avoided overtime based on these quoted times. Thus, the associated plans tended to incur more resource usage costs than overtime costs in comparison to the model based heuristic. However, since this decision to use more resources was made without explicit consideration of overtime costs and errors in the duration estimate provided by the surgeons, this often led to greater total costs. In sum, the model based heuristic outperforms current practice due to better prediction and a more effective scheduling policy. The proportion of the gains due to each of these aspects are analyzed and summarized in the Appendix.

Generally, in a stochastic decision problem, it is not valid to judge the quality of a decision based on an outcome, as due to randomness a good outcome does not necessarily imply a good decision. However, in this work, since the evaluation and validation of the model based heuristic have been extensive, we were confident that they would perform well in the real application. In the final analysis, the real measure of performance of this heuristic is the quality of the decision based on its solution, a question we consider next in the application.

2.6 Application

2.6.1 Implementation

We have implemented the model based heuristic as a decision support system at the operating services department of the UCLA Ronald Reagan Medical Center. Details of this system is provided in the Appendix. The results before and after the implementation across key operational metrics and costs are summarized in Table 2.10. This table shows after implementation, the average number of anesthesiologists on call decreased by 6.7% and average overtime hours for the anesthesiologists on regular duty reduced by 3.7%. This contributed to an increase of average daily utilization across the anesthesiologists by 3.5%. Similarly, the average number of operations rooms used decreased by 8.6% and the average overtime hours at the operating rooms was reduced by 2.7%. This led to an increased average daily utilization across the operating rooms by 3.8%. The improvements in these operational metrics reduced average daily operating room costs by 8.6%, average daily overtime costs by 2.7% and average daily call costs by 8.5%. This translates to an overall average daily cost saving of 7% or estimated to be $2.2 million on an annual basis.

Table 2.10: Summary of Results Before and After Implementation of Decision Support System

Attributes	Before	After	% Change
Average number of anesthesiologists on call per day	6.0	5.6	6.7
Average overtime per day for anesthesiologists (hours)	18.2	17.5	3.7
Average daily utilization of anesthesiologists (%)	75	77.6	3.5
Average number of operating rooms used per day	20.4	18.6	8.6
Average overtime per day for operating rooms (hours)	18.5	18	2.7
Average utilization of operating rooms per day (%)	78	81	3.8
Average daily operating room costs ($)	57,350	52,417	8.6
Average daily overtime costs ($)	2,2375	2,1754	2.8
Average daily call costs ($)	7,145	6,527	8.5
Average total daily costs ($)	86,870	80,729	7.1

The model based heuristic improved upon decision making at operating services due to two main reasons. First, it was more effective at utilizing the flexibility of resources. Most anesthesiologists and operating rooms can perform more than one specialty, typically a primary and a secondary specialty. The model identified these operating room/anesthesiologist combinations and allocated surgeries across these different specialties to them. This led to better usage of resources than the previous approach in which surgeries from a single specialty were assigned to an operating room and anesthesiologist as much as possible. A surgery of a different specialty was assigned to an operating room only when there were high volume of surgeries in a particular day and this was often done without explicit consideration of the allotted anesthesiologists' specialty. Thus, this often required a separate anesthesiologist to perform these surgeries, who were often assigned from call and this was more costly. Second, the model based heuristic explicitly considered uncertainty in surgical durations while determining the daily schedule of an operating room. By using the estimation module, it determined which surgeries could be longer and more uncertain and which surgeries could be shorter and more certain. It then combined long uncertain surgeries with short certain surgeries to effectively utilize gaps in the schedule in each operating room. This in turn reduced the number of operating rooms each day with the resulting cost reduction being more than any potential increases in overtime costs, thus reducing total costs. In contrast, the previous approach used surgeons' predictions of surgery durations. To compensate for the errors in these predictions, planners often underutilized operating rooms by leaving sufficient gaps between surgeries as they did not want to create delays from scheduled start times of succeeding surgeries and incur overtime costs. However, this often led to a larger number of operating rooms being used each day and consequently higher total costs.

Finally, we considered the impact of the schedules generated by our approach on the surgeons. While surgeons were not part of the operating services department, they are a critical element in the system. First, we computed the average idle time between surgeries and found that it reduced by 8 minutes after our work. The surgeons did not find this reduction significant enough to be disruptive, and in fact some of them preferred this as it made their schedule more efficient. Second, we calculated the average number of surgeons per OR per day. Prior to our work, on an average there were 1.54 surgeons per OR per day. After implementing the decision support system, there were 1.57 surgeons per OR per day. This marginal increase suggests that most of the benefits of our approach come from making the correct assignment of operating rooms and anesthesiologist to surgeries and not from increasing the number of surgeons per OR per day. Both these aspects were important to verify that the surgeons were not inconvenienced by the model based approach.

2.6.2 Managerial Insights

We used the model based heuristic to generate several managerial insights. First, we considered the impact of reducing variability in surgical durations. In practice, this could be achieved by better procedures such as check lists, improved information technology, following the correct sequence in tasks and standardized operating protocols derived from best practices. These measures have been advocated by surgeons (Bates and Gawande, 2003; Haynes et al., 2009; Gawande, 2010). In addition, variance can be reduced by improving the prediction of surgical durations. This would require dividing the surgical process in to a series of steps (such as time to incision, skin to skin and closure to exit) and predicting each segment individually as different patient characteristics affect each segment differently (Hosseini et al., 2014). The accuracy of this prediction can be improved by collecting more data on patient characteristics and surgeon experience (Kougias et al., 2012). To consider the impact of variance reduction, we started with the current level of variability in surgical durations and systematically reduced the variance of the distribution of surgical durations across all surgeries by a fixed value. We used these modified distributions to simulate realizations of surgical durations. We then used these data sets to solve the IARSP using the model based heuristic and calculated the resulting costs. These results are summarized by Figure 2.4. This figure shows that the benefits of further reduction in variability decreases and that there is significant diminishing returns on reduction of variability. This suggests that rather than invest in capital-intensive medical equipment to achieve radical reductions in variability in surgical durations, the major cost benefits can be gained by focusing on incremental reduction in variability that can be potentially achieved by better procedures and more detailed data collection for improved predictive analytics.

Figure 2.4: Effect of reducing variability of surgical durations on costs

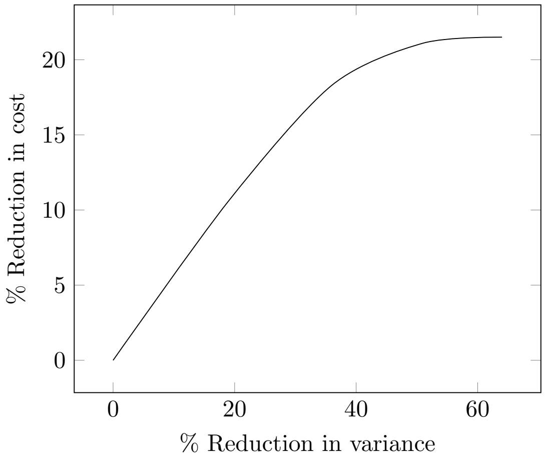
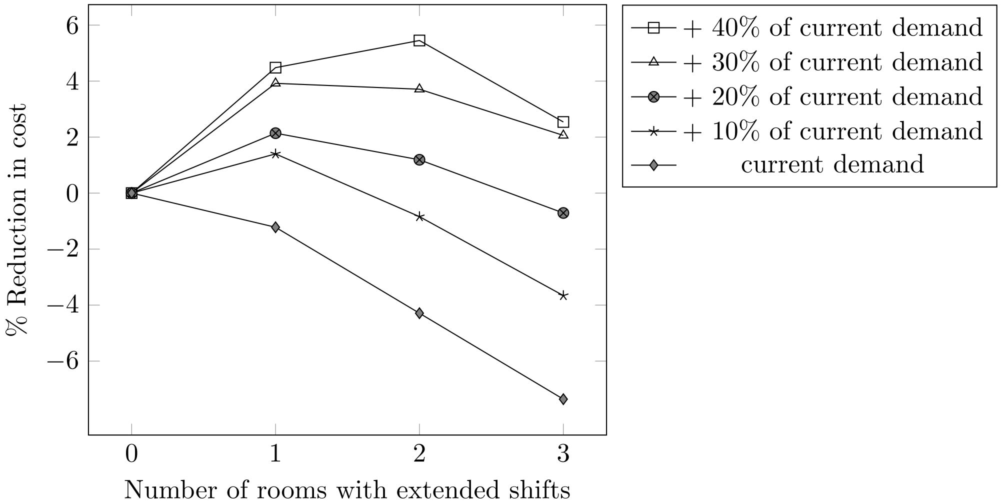


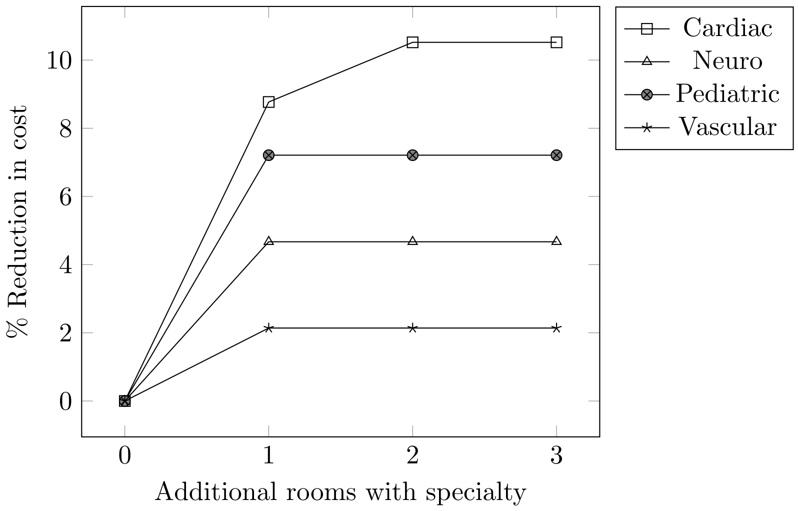
Figure 2.5: Effect of extending shift timings of ORs on costs



Second, we consider the impact of allowing surgeries to start in operating rooms after 3 pm but before the end of the late shift of the anesthesiologists at 7 pm. This would require additional fixed technician and nurse staffing costs. Such extensions can be considered if surgical demand on any day is significantly larger than average daily surgical demand. To perform this analysis, we considered four levels of demand corresponding to increases from daily average demand in surgeries that could occur during the days of any given week. For these scenarios, we incrementally increased the number of operating rooms available after 3 pm by one unit. We then calculated the resulting change in costs from the case when we do not start surgeries after 3 pm but only use the day shift with additional rooms to accommodate such increases in demand. These results summarized in Figure 2.5 suggest that it is beneficial to allow such extensions and the number of operating rooms used depends on the level of demand. This analysis helps management understand how best to react to different levels of daily surgical demand and also estimate the corresponding changes in costs.

Finally, we examined the benefit of increasing cross functionality of the operating rooms. To do so, we considered the various specialties and calculated the potential reduction in costs if the number of operating rooms available for each specialty described in Table 1 is increased. In practice, such increases can be achieved by investing in special equipment to convert general surgery operating rooms to have the cross functionality to accommodate a particular specialty. These results described in Figure 2.6 show that as we increase the number of operating rooms that could be used for a particular specialty, this can lead to a significant reduction in costs and these benefits are often more pronounced in certain specialties. This analysis forms a basis to identify such specialties and determine the priority in which these operating rooms should be made cross functional to enable these additional rooms for the specialties. Further, an additional advantage of making operating rooms cross functional were that a higher number of daily surgeries could be more effectively accommodated without conducting new surgeries after 3 pm. In particular, we found this approach led to at least an additional 5% reduction from the lowest costs attainable for all the demand scenarios considered in Figure 2.5. This provides further justification for management to make the operating rooms more cross functional.

Figure 2.6: Effect of increasing number of specialty rooms on costs



2.6.3 Qualitative Impact

The organizational impact of our work has been significant. Prior to our work, simple rules were used to make important decisions on allocation of anesthesiologists and rooms to surgeries and determining surgery start times. These rules developed based on experience and anecdotal evidence worked well during holidays and weekends when the number of surgeries conducted were low. However, as shown in Section 5 and observed during the implementation, our model significantly outperformed current practice during other days where the number of surgeries performed was high and this resulted in considerable cost savings. Thus, our work demonstrated the value of model based approach and operations research methods in dealing with complexity. This has encouraged the management to investigate other problems in this department using a structured and rigorous approach by employing operations research-based methodologies.

The managerial insights generated from our model have also contributed to the organizational impact. While the effect of variance reduction on improved clinical outcomes has been extensively documented (Neuhauser et al., 2011), our analysis showed that this could also reduce costs. This provided management with the further impetus to implement six sigma programs (Cima et al., 2011) to reduce variability at this department. In addition, our analysis provides management with clear guidance on when to start new surgeries after the day shift and in how many rooms. This provides them with a practical approach to mitigate the impact of varying levels of daily surgical demand on costs and is currently under consideration for implementation in the short term. Finally, we showed the benefits of making some operational rooms cross functional and how to prioritize implementation among the specialties. Furthermore, we demonstrate that this could potentially be a very effective way to accommodate changes in daily surgical demand. While the management at the operating services department was intrigued by this analysis, they felt that there could be significant investments required and this could also lead to disruptions in the schedule while some operating rooms were being reconfigured. Therefore, they are considering this initiative as part of the next broader hospital renovation project.

2.6.4 Limitations

This work has the following limitations. First, the estimation of uncertainty sets can be improved with additional data on the duration of each step in a surgery. However, this data was not available in our application. Second, we do not explicitly consider requests from surgeons for particular start times on a given day and for specific anesthesiologists. While these aspects can be easily incorporated in our model, the management felt that accommodating these requests explicitly can make the overall schedule inefficient and could create additional costs. Therefore, they preferred to make changes to the output in the decision support system only in the most exceptional circumstances. Third, we assume that the overtime payment sufficiently compensates staff for extended shifts. However, in practice, such extensions are unpredictable and staff may not prefer such type of overtime. Thus, there is an implicit inconvenience cost associated with the overtime cost that is not considered in our work. Similarly, we do not consider the inconvenience costs associated with an anesthesiologist being in call, but not being asked to come in to work. While these aspects can be included in our model by suitably appending the overtime and call costs with the appropriate inconvenience costs, quantifying these costs would be challenging. In this regard, recent research in structural estimation (Olivares et al., 2008) could potentially be used to calculate these inconvenience costs and further enhance the outputs of the model. Finally, some anesthesiologists can be used across multiple specialties and this feature was incorporated in our model. However, we do not consider their preferences across specialties as such data was unavailable to us. Future work could focus on all these aspects to improve the model and its ability to attend to the interests of the surgical teams.

In conclusion, the methodology described in this paper has had a major economic and organizational impact at the operating service department at the UCLA RRMC. This organization expects to maintain the described gains and to increase them continuously several years into the future.

2.7 Appendix

Proof of Proposition 1

By definition, the *IARSP* is said to have relatively complete recourse if there exists a feasible second stage solution for any first stage feasible solution (Thiele et al., 2009). Thus, we need to show that there exist feasible $S_i, Over_a, Over_r \forall i \in I, a \in A, r \in R$ satisfying (2.25) through (2.29) $\forall \mathbf{d} \in \boldsymbol{\mathcal{D}}(\tau)$ $\forall (\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}) \in K_1$, where K_1 is the feasibility set of the first stage problem.

Let $\bar{S}_i = \max_{a \in A} t_a^{start} + \sum_{j \in I - \{i\}} \left(\bar{d}_j + \hat{d}_j \right)$ define an upper bound on S_i, the start time of surgery i. This follows as $\bar{S}_i$ would be the start time of surgery i if all surgeries start after the start time of the last shift ($\max_{a \in A} t_a^{start}$), i is the last surgery to be performed in the day after all the other surgeries have been performed and every other surgery j lasts for its maximum allowed duration $\left(\bar{d}_j + \hat{d}_j \right)$. Thus, S_i is bounded from above. We define $M_{seq} = \max_{i \in I} \bar{S}_i + \max_{i \in I} \left(\bar{d}_j + \hat{d}_j \right)$.

Let $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}) \in K_1$ be some first stage feasible solution. We next show that $S_i = \max_{h \in I} \{s_h\} + \sum_{h \in I, h \neq i} u_{hi} d_h \ \forall i \in I$ is a feasible solution to (2.25). First observe that by our definition of M_{seq}, this constraint is deactivated when $u_{ij} = 0$. To check the feasibility when $u_{ij} = 1$ note from (2.12) that $u_{ji} = 0$. Substituting the values of S_i and S_i in (2.25), we get:

$$\max_{h \in I} \{s_h\} + \sum_{\substack{h \in I \\ h \neq j}} u_{hj} d_h \geq \max_{h \in I} \{s_h\} + \sum_{\substack{h \in I \\ h \neq i}} u_{hi} d_h + d_i - M_{seq} (1 - u_{ij}) \quad \forall i, j \in I \tag{2.47}$$

Separating the term $u_{ij} d_i$ from the summation on the LHS and the term $u_{ji} d_j$

from the summation on the RHS leads to:

$$\max_{h\in I}\{s_h\} + \sum_{\substack{h\in I\\ h\neq i,j}} u_{hj}d_h + u_{ij}d_i \geq \max_{h\in I}\{s_h\} + \sum_{\substack{h\in I\\ h\neq i,j}} u_{hi}d_h + d_i + u_{ji}d_j - M_{seq}(1-u_{ij}) \qquad \forall i,j\in I \tag{2.48}$$

Next, setting $u_{ij} = 1, u_{ji} = 0$ and simplifying the above expression we get,

$$\sum_{\substack{h\in I\\ h\neq i,j}} d_h (u_{hj} - u_{hi}) \geq 0 \tag{2.49}$$

From (**??**) $u_{hj} \geq u_{hi} + u_{ij} - 1$. As $u_{ij} = 1$, this implies, $u_{hj} - u_{hi} \geq 0$. Therefore, $S_i = \max_{h\in I}\{s_h\} + \sum_{h\in I, h\neq i} u_{hi}d_h \ \forall i \in I$ is a feasible solution to (2.25) $\forall(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) \in K_1$. Note that (2.26) is satisfied since $S_i \geq s_i$. Since there exists a feasible $S_i\ \forall i \in I$ in the recourse problem, then there will always exist $Over_a \geq 0\ \forall a \in A$ and $Over_r \geq 0\ \forall r \in R$ for any given $\mathbf{x},\mathbf{z}$ satisfying (2.27) through (2.29). Thus, the *IARSP* has relatively complete recourse. This result can be extended to the case where the integrality condition on $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ is relaxed. The proof of this extension is provided in 2.7.

Proof of Proposition 2

For a given first stage solution $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ and an arbitrary $\mathbf{d} \in \boldsymbol{\mathcal{D}}(\tau)$, the dual of the recourse function $\boldsymbol{\mathcal{R}}^{\mathbf{D}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d})$ is given by:

$$\begin{aligned}\boldsymbol{\mathcal{R}}^{\mathbf{D}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d}) = \max \Bigg\{ & \sum_{i\in I} d_i \left(\sum_{j\in I-\{i\}} \lambda_{ij} + \sum_{a\in A} \mu_{ia} + \sum_{r\in R} \theta_{ir} \right) \\ & + \sum_{i\in I} s_i\phi_i - M_{seq} \sum_{i,j\in I i\neq j} \lambda_{ij}\left(1-u_{ij}\right) \\ & - M_{anesth} \sum_{\substack{i\in I\\ a\in A}} \mu_{ia}\left(1 - x_{ia} + y_a\right) - M_{room} \sum_{\substack{i\in I\\ r\in R}} \theta_{ir}\left(1 - z_{ir}\right) \\ & - \sum_{\substack{i\in I\\ r\in R}} \theta_{ir} T^{end} - \sum_{\substack{i\in I\\ a\in A}} \mu_{ia} t_a^{end} \Bigg\}\end{aligned}$$

subject to,

$$\sum_{i\in I} \mu_{ia} \leq c_{oa} \qquad \forall a \in A \qquad (2.50)$$

$$\sum_{i\in I} \theta_{ir} \leq c_{or} \qquad \forall r \in R \qquad (2.51)$$

$$\sum_{\substack{j\in I\\ j\neq i}} \lambda_{ij} - \sum_{\substack{j\in I\\ j\neq i}} \lambda_{ji} + \sum_{a\in A} \mu_{ia} + \sum_{r\in R} \theta_{ir} - \phi_i \geq 0 \qquad \forall i \in I \qquad (2.52)$$

$$\lambda_{ij}, \mu_{ia}, \theta_{ir}, \phi_i \geq 0 \qquad \forall i, j \in I, a \in A, r \in R \qquad (2.53)$$

Here, $\lambda_{ij},\ \mu_{ia},\ \theta_{ir},\ \phi_i\ \forall i, j \in I, a \in A, r \in R$ are dual variables corresponding to constraints (2.25)-(2.29) respectively. $\boldsymbol{\mathcal{Q}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s})$ can now be written as

$$\begin{aligned}\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max \Bigg\{ & \sum_{i\in I}\left(\bar{d}_i + f_i\hat{d}_i\right)\left(\sum_{j\in I-\{i\}}\lambda_{ij} + \sum_{a\in A}\mu_{ia} + \sum_{r\in R}\theta_{ir}\right) \\ & + \sum_{i\in I} s_i\phi_i - M_{seq}\sum_{i,j\in I i\neq j}\lambda_{ij}\left(1-u_{ij}\right) \\ & - M_{anesth}\sum_{\substack{i\in I\\a\in A}}\mu_{ia}\left(1-x_{ia}+y_a\right) - M_{room}\sum_{\substack{i\in I\\r\in R}}\theta_{ir}\left(1-z_{ir}\right) \\ & - \sum_{\substack{i\in I\\r\in R}}\theta_{ir}T^{end} - \sum_{\substack{i\in I\\a\in A}}\mu_{ia}t_a^{end}\Bigg\}\end{aligned}$$

subject to,

$$(2.50) - (2.53)$$

$$\sum_{i\in I}|f_i| \leq \tau \tag{2.54}$$

$$-1 \leq f_i \leq 1 \qquad \forall i \in I \tag{2.55}$$

Simplifying the objective function,

$$\begin{aligned}\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max \Bigg\{ & \sum_{i\in I}\left(\bar{d}_i\pi_i + f_i\pi_i\hat{d}_i\right) + \sum_{i\in I} s_i\phi_i \\ & - M_{seq}\sum_{i,j\in I i\neq j}\lambda_{ij}\left(1-u_{ij}\right) - M_{anesth}\sum_{\substack{i\in I\\a\in A}}\mu_{ia}\left(1-x_{ia}+y_a\right) \\ & - M_{room}\sum_{\substack{i\in I\\r\in R}}\theta_{ir}\left(1-z_{ir}\right) - \sum_{\substack{i\in I\\r\in R}}\theta_{ir}T^{end} - \sum_{\substack{i\in I\\a\in A}}\mu_{ia}t_a^{end}\Bigg\}\end{aligned}$$

subject to, $(2.50) - (2.55)$

$$\pi_i = \sum_{j\in I-\{i\}}\lambda_{ij} + \sum_{a\in A}\mu_{ia} + \sum_{r\in R}\theta_{ir} \qquad \forall i \in I \tag{2.56}$$

$$\pi_i \geq 0 \qquad \forall i \in I \tag{2.57}$$

$\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ is a bilinear optimization problem because of the presence of bilinear terms $f_i\pi_i$ in the objective function. Further, at optimality $f_i^* \geq 0 \; \forall i \in I$. This is true as $\pi_i \geq 0$ and the feasibility set of f_i is independent of π_i. If τ is a positive integer, constraints (2.54) and (2.55) constrain the feasibility set of f_i $\forall i \in I$ to be the set $\{0, 1\}$. This implies that:

$$f_i\pi_i = \begin{cases} \pi_i, & \text{if } f_i = 1 \\ 0, & \text{if } f_i = 0 \end{cases} \qquad \forall i \in I$$

We introduce an additional variable $\xi_i = f_i\pi_i$ and rewrite $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ as follows:

$$\begin{aligned} \boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max \Bigg\{ & \sum_{i\in I}\left(\bar{d}_i\pi_i + \xi_i\hat{d}_i\right) - M_{seq}\sum_{i,j\in I i\neq j}\lambda_{ij}\left(1-u_{ij}\right) + \sum_{i\in I} s_i\phi_i \\ & - M_{anesth}\sum_{\substack{i\in I\\ a\in A}}\mu_{ia}\left(1 - x_{ia} + y_a\right) - M_{room}\sum_{\substack{i\in I\\ r\in R}}\theta_{ir}\left(1 - z_{ir}\right) \\ & - \sum_{\substack{i\in I\\ r\in R}}\theta_{ir}T^{end} - \sum_{\substack{i\in I\\ a\in A}}\mu_{ia}t_a^{end} \Bigg\} \end{aligned}$$

subject to, $\quad (2.51) - (2.57)$

$$\xi_i \leq M_f f_i \qquad \forall i \in I \qquad (2.58)$$

$$\xi_i \leq \pi_i \qquad \forall i \in I \qquad (2.59)$$

$$\xi_i \geq 0 \qquad \forall i \in I \qquad (2.60)$$

Where, M_f is a sufficiently large positive number. Thus, $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ can be written as a mixed-integer program.

Proof of Proposition 3

This proof is adapted from Thiele et al. (2009). For conciseness of notation we represent the set defined by (2.50)-(2.53) as $\Lambda(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$. Also, let $\boldsymbol{\lambda} = (\lambda_{ij})\forall i,j \in I$, $\boldsymbol{\mu} = \mu_{ia}\forall i \in I, a \in A$, $\boldsymbol{\theta} = \theta_{ir}\forall i \in I, r \in R$, $\boldsymbol{\phi} = \phi_i\forall i \in I$.

From the proof of Proposition 2 and strong duality, we get:

$$\begin{aligned}\boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}) = \max_{\substack{(\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\theta},\boldsymbol{\phi})\\ \in \\ \Lambda(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})}} & \left\{ \sum_{i\in I} d_i \left(\sum_{j\in I-\{i\}} \lambda_{ij} + \sum_{a\in A} \mu_{ia} + \sum_{r\in R} \theta_{ir} \right) \right. \\ & + \sum_{i\in I} s_i\phi_i - M_{seq} \sum_{i,j\in I i\neq j} \lambda_{ij}\left(1-u_{ij}\right) \\ & - M_{anesth} \sum_{\substack{i\in I\\ a\in A}} \mu_{ia}\left(1 - x_{ia} + y_a\right) \\ & \left. - M_{room} \sum_{\substack{i\in I\\ r\in R}} \theta_{ir}\left(1-z_{ir}\right) - \sum_{\substack{i\in I\\ r\in R}} \theta_{ir}T^{end} - \sum_{\substack{i\in I\\ a\in A}} \mu_{ia} t_a^{end} \right\}\end{aligned} \tag{2.61}$$

Let $\boldsymbol{d}^l \in \arg\max_{\boldsymbol{d}\in\mathcal{D}(\tau)} \boldsymbol{\mathcal{R}}(\mathbf{x}^l,\mathbf{y}^l,\mathbf{z}^l,\mathbf{u}^l,\mathbf{s}^l,\mathbf{d})$. From equation (2.21), $\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) = \max_{\mathbf{d}\in\mathcal{D}(\tau)} \boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$. Then for an arbitrary $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$,

$$\boldsymbol{\mathcal{Q}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) \geq \boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}^l) \tag{2.62}$$

Furthermore, let $(\boldsymbol{\lambda}^l,\boldsymbol{\mu}^l,\boldsymbol{\theta}^l,\boldsymbol{\phi}^l)$ be an optimal solution of (2.61) with $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}) = (\mathbf{x}^l,\mathbf{y}^l,\mathbf{z}^l,\mathbf{u}^l,\mathbf{s}^l,\mathbf{d}^l)$. From (2.61) we get:

$$\begin{aligned}\boldsymbol{\mathcal{R}}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d}^l) \geq & \sum_{i\in I} d_i^l \left(\sum_{j\in I-\{i\}} \lambda_{ij}^l + \sum_{a\in A} \mu_{ia}^l + \sum_{r\in R} \theta_{ir}^l \right) + \sum_{i\in I} s_i\phi_i^l \\ & - M_{seq} \sum_{i,j\in I i\neq j} \lambda_{ij}^l\left(1-u_{ij}\right) \\ & - M_{anesth} \sum_{\substack{i\in I\\ a\in A}} \mu_{ia}^l\left(1 - x_{ia} + y_a\right) - M_{room} \sum_{\substack{i\in I\\ r\in R}} \theta_{ir}^l\left(1-z_{ir}\right) \\ & - \sum_{\substack{i\in I\\ r\in R}} \theta_{ir}^l T^{end} - \sum_{\substack{i\in I\\ a\in A}} \mu_{ia}^l t_a^{end}\end{aligned} \tag{2.63}$$

From, equation (2.62) and (2.63) we get,

$$\begin{aligned}\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{s}) \geq & \sum_{i \in I} d_i^l \left(\sum_{j \in I-\{i\}} \lambda_{ij}^l + \sum_{a \in A} \mu_{ia}^l + \sum_{r \in R} \theta_{ir}^l \right) + \sum_{i \in I} s_i \phi_i^l \\ & - M_{seq} \sum_{i,j \in I i \neq j} \lambda_{ij}^l \left(1 - u_{ij}\right) \\ & - M_{anesth} \sum_{\substack{i \in I \\ a \in A}} \mu_{ia}^l \left(1 - x_{ia} + y_a\right) - M_{room} \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l \left(1 - z_{ir}\right) \\ & - \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l T^{end} - \sum_{\substack{i \in I \\ a \in A}} \mu_{ia}^l t_a^{end} \qquad (2.64)\end{aligned}$$

From (2.56) and (2.64) we get:

$$\begin{aligned}\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{s}) \geq & \sum_{i \in I} \left(\bar{d}_i \pi_i^l + \xi_i^l \hat{d}_i \right) + \sum_{i \in I} s_i \phi_i^l - M_{seq} \sum_{i,j \in I i \neq j} \lambda_{ij}^l \left(1 - u_{ij}\right) \\ & - M_{anesth} \sum_{\substack{i \in I \\ a \in A}} \mu_{ia}^l \left(1 - x_{ia} + y_a\right) - M_{room} \sum_{\substack{i \in I \\ r \in R}} \theta_{ir}^l \left(1 - z_{ir}\right) \\ & - \sum_{\substack{i \in I \\ r \in R}} \theta_{ir} T^{end} - \sum_{\substack{i \in I \\ a \in A}} \mu_{ia}^l t_a^{end} \qquad (2.65)\end{aligned}$$

Therefore, the right hand side of (2.65) is a lower bound on $\mathcal{Q}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$.

Proof of Proposition 4

The value function of a linear program $z(b) = \min \left\{ c^T x | Ax \geq b \right\}$ is a piecewise linear convex function over the domain for which the linear program is feasible (Martin, 1999)[Corollary 2.49, pp. 75].

From the proof of the extension of Proposition 1 given in 2.7, for any $\mathbf{d} \in \mathcal{D}(\tau)$, if K_1^{LP} is the feasibility set of the first stage problem with the integrality condition on $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ relaxed and $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) \in K_1^{LP}$, $mathbf\mathcal{R}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{d})$ is a feasible linear program over its domain as $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ appear in the right hand side of its constraints. Thus, from above, $\mathcal{R}(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{s})$ is a piecewise linear convex function in $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s},\mathbf{s})$ for any given $\mathbf{d} \in$ $mathcalD(\tau)$. From equation (2.21), $\mathcal{Q}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{s}) = \max_{\mathbf{d} \in \mathcal{D}(\tau)} \mathcal{R}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{s})$. Therefore, $\mathcal{Q}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{s})$ is also piecewise-linear and convex in $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{s})$.

Proof of Proposition 5

This proposition is proved by using an interchange argument. We compare two solutions. The first solution called Solution 1 is when $u_{ij} = 1, u_{ji} = 0$ with recourse variables $S_k, Over_a$ and $Over_r$. The second solution called Solution 2 is when $u_{ij} = 0, u_{ji} = 1$ with recourse variables $S'_k, Over'_a, Over'_r$. Let there be p surgeries between i and j. The first stage variables in the two solutions do not differ for any surgery preceding and following surgeries i, j, k $\forall k \in \{1, 2, \ldots, p\}$. In the recourse of Solution 1, the following would be true,

$$S_k \geq S_i + \bar{d}_i + \hat{d}_i f_i \qquad k \in \{1, 2, ..., p\} \tag{2.66}$$

Similarly, in the recourse of Solution 2, the following would be true,

$$S'_k \geq S'_j + \bar{d}_j + \hat{d}_j f_j \qquad k \in \{1, 2, ..., p\} \tag{2.67}$$

Note that $S'_j = S_i$. If that were not true and $S'_j < S_i$, this would imply that there exists a $\delta > 0$ such that $S'_j = S_i - \delta$ is a feasible solution to Solution 2. Also since surgeries i, j are done by the same anesthesiologists and in the same rooms, and the first stage variables of all the preceding surgeries are not changed in the two solutions, then in Solution 1, $S_i - \delta$ is a feasible start time to surgery i, which would imply $S_k \geq S_i - \delta + \bar{d}_i + \hat{d}_i f_i$. Then the surgeries $k \in \{1, 2, \ldots, p\}$ may start earlier by δ and all surgeries following these surgeries may be brought forward. This would imply that the recourse cost can be further decreased, which would be a contradiction to Solution 1 in which S_i is already the solution of the minimization of the recourse problem. We can use similar logic to prove that $S_i < S'_j$ would also lead to a contradiction to Solution 2. Thus, $S'_j = S_i$.

As $\bar{d}_j \geq \bar{d}_i$, $\hat{d}_j \geq \hat{d}_i$ and $S'_j = S_i$ this implies that $S'_k \geq S_k$ for $k \in \{1, 2, ..., p\}$. This also implies that $S'_h \geq S_h$ $\forall h \in I$ such that $u_{kh} = 1$ since from (28), $S'_h \geq S'_k + \bar{d}_k + \hat{d}_k f_k - M_{seq}(1 - u_{kh}) \geq S_k + \bar{d}_k + \hat{d}_k f_k - M_{seq}(1 - u_{kh}) = S_h$ $\forall k, h \in I$

Hence, from (2.27) and (2.28), $Over'_r \geq Over_r$ and $Over'_a \geq Over_a$. Thus, for any surgery duration, the objective function with the longer surgery first would have higher overtime.

The Assignment and Scheduling Decision Making Environment

Assignment and scheduling decisions at operating services department in the UCLA RRMC is complicated by the large number of Operating Rooms (ORs) and anesthesiologists, variety in surgical procedures, variability in anesthesiologist workload and unpredictability in surgical durations. This section provides more details on these aspects. Table 2.11 shows the average number of procedures performed per day and the range in the number of surgeries across the various specialties. Figure 2.7 shows the uncertainty in the total anesthesiologist workload per day. The two most apparent sources of this variability are in the number of surgeries and the mix of specialties. However, Figure 2.8 shows that even if we control for these factors, there is still considerable variability in total workload per day. This variability arises because of differences across procedures within a specialty and because different patients react differently even to the same procedure (Schaefer et al., 2005). As a direct consequence, the required number of resources (i.e., rooms with anesthesiologists) at any instant as shown in Figure 2.9 and surgical durations are unpredictable. Specifically, the coefficient of variation of surgery duration across these 2700 types procedures varies from 0.75% to 125%.

Sample Average Approximation Solution for the Integrated Anesthesiologist and Room Scheduling Problem

In this section we provide the formulation and a short description of sample average approximation procedure based solution to solve the integrated anesthesiologist and room scheduling problem.

The variable and parameter description is as described in §2.3 in the paper. The difference between the sample average approximation formulation and the robust optimization formulation is that the surgery duration d_i $\forall i \in I$, instead of belonging to an uncertainty set are now modeled as random variables and are denoted by $d_i(\omega)$ under scenario $\omega \in \Omega$. For $|I|$ surgeries the random vector $\mathbf{d}(\omega) = \left\{d_1(\omega), d_2(\omega), \ldots, d_{|I|}(\omega)\right\}$ is the vector of surgery durations under scenario $\omega \in \Omega$. The support of $\mathbf{d}$ is $\mathbb{R}_+^{|I|}$. Following the evidence found in clinical literature (Strum et al., 2000) and standard assumptions in surgery scheduling literature (Batun et al., 2011; Denton et al., 2010) we assumed that these d_i are independent and have log-normal distribution. The standard formulation of the resulting two-stage stochastic program with recourse is:

$$\min \left\{ \sum_{r \in R} c_r v_r + \sum_{a \in A} c_q y_a + \mathbb{E}\left[\boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d}(\omega))\right] \right\} \tag{2.68}$$

subject to,

$$(2.2) - (2.20)$$

and,

$$\boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d}(\omega)) = \min \left\{ \sum_{a \in A} c_{oa} Over_a(\omega) + \sum_{r \in R} c_{or} Over_r(\omega) \right\} \tag{2.69}$$

subject to,

$$S_j(\omega) \geq S_i(\omega) + d_i(\omega) - M_{seq}(1 - u_{ij}) \; \forall i, j \in I \tag{2.70}$$

$$S_i(\omega) \geq s_i \; \forall i \in I \tag{2.71}$$

$$\begin{aligned} Over_a(\omega) \geq{}& S_i(\omega) + d_i(\omega) - t_a^{end} \\ & - M_{anesth}(1 - x_{ia} + y_a) \; \forall i \in I, a \in A \end{aligned} \tag{2.72}$$

$$Over_r(\omega) \geq S_i(\omega) + d_i(\omega) - T^{end} - M_{room}(1 - z_{ir}) \; \forall i \in I, r \in R \tag{2.73}$$

$$S_i(\omega), Over_a(\omega), Over_r(\omega) \geq 0; \forall i \in I, a \in A, r \in R \tag{2.74}$$

Assuming a log-normal distribution we estimate the conditional mean and standard deviation for each cluster of CPT codes. From this estimated distribution we draw $N_s = 1000$ samples and formulate the sample average approximation of the two-stage stochastic program as:

$$\min \left\{ \sum_{r \in R} c_r + \sum_{a \in A} c_q y_a + \frac{1}{N_s} \sum_{n=1}^{N_s} \boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d}^n) \right\} \tag{2.75}$$

subject to,

$$(2.2) - (2.20)$$

and,

$$\boldsymbol{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}, \mathbf{d}^n) = \min \left\{ \sum_{a \in A} c_{oa} Over_a^n + \sum_{r \in R} c_{or} Over_r^n \right\} \tag{2.76}$$

subject to,

$$S_j^n \geq S_i^n + d_i^n - M_{seq}(1 - u_{ij}) \qquad \forall i, j \in I \tag{2.77}$$

$$S_i^n \geq s_i \qquad \forall i \in I \tag{2.78}$$

$$Over_a^n \geq S_i^n + d_i^n - t_a^{end} - M_{anesth}(1 - x_{ia} + y_a) \qquad \forall i \in I, a \in A \tag{2.79}$$

$$Over_r^n \geq S_i^n + d_i^n - T^{end} - M_{room}(1 - z_{ir}) \qquad \forall i \in I, r \in R \tag{2.80}$$

$$S_i^n, Over_a^n, Over_r^n \geq 0 \qquad \forall i \in I, a \in A, r \in R \tag{2.81}$$

We solve the above program by the integer L-shaped decomposition framework (Birge and Louveaux, 2011). The results of the solution procedure are provided in Table 2.7 of the main paper.

Quantile Regression Procedure for Estimating Uncertainty Set $\boldsymbol{\mathcal{D}}(\tau)$

We develop a quantile regression model for predicting $g_L(\widetilde{\mathbf{b};}\rho)$ and $g_U(\widetilde{\mathbf{b};}\rho)$ which are used in (2.43) to estimate the uncertainty set $\boldsymbol{\mathcal{D}}(\tau)$. As the objective of this model was to predict quantiles, we prefer a parsimonious model with predictive power rather than an over-specified model. For evaluation of the predictive power of this model we use an out of sample Mean Square Prediction Error (MSPE) with respect to the conditional median. To avoid over-specification of the model, we use an Akaike Information Criterion (AIC). The dataset Δ^E had 25700 observations. Δ^E was divided into two disjoint datasets $\Delta^{E-Train}$ (19300 observations) and Δ^{E-Test} (6400 observations). The quantile regression was built on $\Delta^{E-Train}$ and the out-of-sample MSPE was performed on Δ^{E-Test}. The regression model was built using the following steps:

Step 1: Dimension Reduction. Due to the large number of CPT codes and surgeons compared to the number of observations, we performed clustering to reduce the number of factor variables corresponding to CPT codes and surgeons. For surgeries with multiple CPT codes, we concatenated the CPT codes to create a composite code. For example, if a procedure consisted of CPT codes *A, B* and *C*, we created the composite code *A-B-C*. After this procedure the total number of unique surgery types grew from 2706 to 4061 as there are now more unique procedures with each combination being a different code. Subsequently, we performed a *k-means* clustering on the median values of observed surgery duration for each unique procedure code. This clustering procedure is similar to that performed in He et al. (2012). We plot the number of clusters against the % variance explained and choose the elbow point for selecting the total number of clusters. We perform similar clustering procedure for the surgeons. From Figure 2.10 and 2.11 we choose 6 clusters for both CPT codes and surgeons. This choice of clusters explained 95% of the variance in median surgical durations. We use these clustered variables in the subsequent quantile regression and name them $CPTCLUST$ for the CPT clusters and $PROVCLUST$ for the surgeons.

Step 2: Quantile Regression From the dataset Δ^E, there are 8 possible explanatory variables that could be included for modeling surgical durations. However, using the CPT cluster variable ($CPTCLUST$) and surgery service variable (*SERVICE*) at the same time led to an ill-defined model matrix. This was due to the collinearity between these two variables as measured by the Variance Inflation Factor (VIF) as described in Hair et al. (2006). The VIF for *SERVICE* was 40.19. We removed this variable and found that none of the other variables demonstrated a $VIF > 10$, thus meeting the criteria for not exhibiting significant collinearity (Table 2.12).

Subsequently, we performed quantile regression incorporating the above 7 variables. We then sequentially remove variables in increasing order of significance and compare the out-of-sample MSPE and the AIC for the resulting models. The results are shown in Table 2.13. Incorporating surgeon information (i.e. $PROVCLUST$), CPT information (i.e., $CPTCLUST$), age and ASA score improves the prediction and leads to a lower AIC value. Therefore, these variables were incorporated in the model. The number of CPT codes (NUMCPT) was excluded from the final model specification as adding it did not lead to any improvement in MSPE or AIC. The final specification of the quantile regression model was thus,

$$\begin{aligned} ACTUALHRS = {} & BOOKEDHRS + ASA + AGE + PATCLASS \\ & + CPTCLUST + PROVCLUST \end{aligned} \tag{2.82}$$

Next, in Table 2.14 we provide the coefficients of the above quantile regression model at several quantile values (10^{th}, 30^{th}, 60^{th} and 90^{th} quantile). This table demonstrates an advantage of using quantile regression instead of OLS regression. In particular, it allows for the effect of variable to be different at different quantiles. The coefficient of BOOKEDHRS, representing the surgeons' estimate of surgery duration, increases with increasing quantiles of surgery duration. This implies that more weight is provided to surgeons estimate for longer surgeries. This could imply that surgeons provide a more accurate estimate for longer surgeries than for shorter surgeries. One reason for this could be that surgeons often round up or down to the nearest quarter of an hour while providing estimates. This rounding off would lead to more significant differences for shorter surgeries than for longer surgeries. Another variable that demonstrates changing coefficients with quantiles is ASA. The coefficient of the ASA code increases with increasing quantiles. This could imply that for longer surgeries each increase in ASA score contributes more to the surgery duration than for shorter surgeries. This is also intuitive as the fitness of the patient before surgery would be more significant for longer and more complex procedures.

The Decision Support System at the UCLA RRMC

The decision support system for integrated anesthesiologist and operating room scheduling at the UCLA RRMC was built using the Python programming language and a schematic of this system is shown in Figure 2.12. This system consists of an uncertainty set estimation module and an optimization module. The surgery characteristics were provided using the CareConnect database at the hospital and inputted to the uncertainty set estimation module. The output from this module was sent along with anesthesiologist availability and the surgery resource specialty information provided by the Qgenda database to the optimization module. This solved the IARSP using the model based heuristic and generated an optimized schedule for the following day specifying the assignment decisions of anesthesiologists and operating room to surgeries along with their scheduled start times. The output of this module was provided to the planner who made adjustments as needed to accommodate special requests by surgeons to change start times or for rooms with additional specialized equipment or for specific anesthesiologists.

There were several challenges in implementing this system. First, the CareConnect database and the Qgenda database had to be accessed daily and their output had to be reformatted to be compatible with the uncertainty set estimation and the optimization modules. This necessitated the development of a specialized automated interface which required regular maintenance. Second, the planners were initially skeptical about the ability of our system to consider all specialties and constraints. Further, they were unsure if the prediction of surgical durations was better than estimates made by the surgeons and if the model had adequately captured uncertainty in surgical durations. They felt that if these aspects were not effectively incorporated, this could lead to schedule disruptions, unsatisfied patients and extensive overtime costs. To ensure that the planners were confident with this system, we ran the model in parallel with their approach as described in §2.5.2 of the paper, so that they could understand the solution of the model and compare this with their own rules. They were reassured that the model solution corresponded to their solution when the number of daily surgeries was small. However, they also appreciated how the model solution outperformed their rules when the number of daily surgeries was high and the resulting scheduling complexity was larger. In this scenario, the model was more effective in utilizing the flexibility of the multiple parallel resources and considering uncertainty in surgical durations. A third challenge in implementing the system was that surgical characteristics had to be updated and new procedures added to the system. Since these required specialized clinical input, a formalized procedure had to be instituted where the feedback of the surgeons and anesthesiologists had to be solicited and manually updated in the CareConnect database. This was time consuming and had to be done in monthly basis. However, this was essential to ensure the continued efficacy of our system.

Evaluation of the Prediction and Scheduling Benefits of the Model Based Heuristic

In this section, we analyze what proportion of the gains were due to a better prediction method for surgical durations and how much was due to the scheduling policy. To conduct this analysis, we used the estimates of surgical durations made by the surgeons and ran our model for the live and historical validation. Note that the resulting gains would now be entirely from the scheduling policy. We then subtracted these gains from the original benefits which considered prediction and scheduling to calculate the gains of the prediction method. This analysis showed that on average 41% of the benefits were due to better prediction and 59% was due a better scheduling policy. These results demonstrate that due to the complexity of the problem, the most benefits can be got by combining prediction and optimization, an aspect for which robust optimization is particularly well suited.

Proof of Extension of Proposition 1

Let $IARSP^{LP}$ be the form of the $IARSP$ with the integrality conditions on $(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s})$ relaxed. Let the feasibility set of $IARSP^{LP}$ be denoted by K_1^{LP}. To prove that $ISRSP^{LP}$ has relatively complete recourse we show by mathematical induction that there exist feasible $S_i \geq 0, Over_a \geq 0, Over_r \geq 0 \; \forall i \in I, a \in A, r \in R \; \forall(\mathbf{x},\mathbf{y},\mathbf{z},\mathbf{u},\mathbf{s}) \in K_1^{LP}$ and $\forall \mathbf{d} \in \boldsymbol{\mathcal{D}}(\tau)$. The proof of Proposition 1 implies that it is now sufficient to prove the existence of a feasible $S_i \forall i \in I$ that satisfies (2.25). We consider a base case and an induction step, to show this result.

Base case: Let there be two surgeries $I = \{1, 2\}$ and also let there be some

$(\mathbf{x}, \mathbf{z}, \mathbf{u}, \mathbf{s}) \in K_1$ and $\mathbf{d} \in \boldsymbol{\mathcal{D}}(\tau)$. Further, let $M_{seq} = \sum_{i\in\{1,2\}}(\bar{d_i} + \hat{d_i})$. Then the following is a feasible solution to (2.25) through (2.29),

$$S_1 = s_1 + s_2 + d_2 + M_{seq}(1 - u_{12}) \tag{2.83}$$

$$S_2 = s_1 + s_2 + M_{seq} \tag{2.84}$$

Proof: Since $u_{12} \leq 1$, (2.26) is satisfied. For these two surgeries, (2.25) can be written as,

$$S_2 \geq S_1 + d_1 - M_{seq}(1 - u_{12}) \tag{2.85}$$

$$S_1 \geq S_2 + d_2 - M_{seq}(1 - u_{21}) \tag{2.86}$$

Substituting the value of S_1 and S_2 from (2.83) and (2.84) in (2.85),

$$s_1 + s_2 + M_{seq} \geq s_1 + s_2 + d_1 + d_2 + M_{seq}(1 - u_{12}) - M_{seq}(1 - u_{12}) \tag{2.87}$$

The above simplifies to $M_{seq} \geq d_i + d_2$, which is true by our choice of M_{seq}. Next, substituting the values of S_1 and S_2 in (2.86),

$$s_1 + s_2 + d_2 + M_{seq}(1 - u_{12}) \geq s_1 + s_2 + M_{seq} + d_2 - M_{seq}(1 - u_{21}) \tag{2.88}$$

This simplifies to $M_{seq} \geq M_{seq}(u_{12} + u_{21})$ which is true since $u_{12} + u_{21} \leq 1$ from (2.12). Therefore, the values of S_1 and S_2 satisfy (2.25) and $IARSP^{LP}$ has relatively complete recourse for $I = \{1, 2\}$.

Induction step: Let $IARSP^{LP}$ have relatively complete recourse for $I = \{1, 2, \ldots, k-1\}$ and for this I, $M_{seq} = M'$ and S'_i $\forall i \in I$ be a feasible solution. Thus,

$$S'_j \geq S'_i + d_i - M'(1 - u_{ij}) \qquad \forall i, j \in \{1, 2, \ldots, k-1\} \tag{2.89}$$

We now prove that $IARSP^{LP}$ has relatively complete recourse for $I = \{1, 2, \ldots, k-1, k\}$.

Proof: Let $I = \{1, 2, \ldots, k\}$, $M_{seq} = 2M'$ and let there be some $(\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{s}) \in K_1^{LP}$ and $\mathbf{d} \in \boldsymbol{\mathcal{D}}(\tau)$.

We show that there exists feasible S_i $\forall i \in I$. To show this, we first provide a feasible solution for S_i $\forall i \in I - \{k\}$. Then we show that for any such feasible S_i $\forall i \in I - \{k\}$, there exists a feasible S_k. We define the following feasible solution for $i \in I - \{k\}$

$$S_i = S'_i + d_k + \max_{h \in I} \{s_h\} + M'(1 - u_{ik}) \qquad \forall i \in I - \{k\} \tag{2.90}$$

All such S_i satisfy (2.26). To verify that they satisfy (2.25), we substitute the values of S_i and M_{seq} in (2.25).

$$\begin{aligned} S'_j + d_k + \max_{h \in I} \{s_h\} + M'(1 - u_{jk}) \geq S'_i + d_k + \max_{h \in I} \{s_h\} + M'(1 - u_{ik}) \\ + d_i - 2M'(1 - u_{ij}) \qquad \forall i, j \in \{1, 2, \ldots, k - 1\} \end{aligned}$$

Simplifying, we get,

$$S'_j \geq S'_i + d_i - M'(1 - u_{ij}) - M'(1 + u_{ik} - u_{jk} - u_{ij}) \tag{2.91}$$

From (2.89), $S'_j \geq S'_i + d_i - M'(1 - u_{ij})$. From (**??**), $1 + u_{ik} - u_{ij} - u_{jk} \geq 0$. Thus, the above is always true. Next we show that for all such S_i given by (2.90) there exists feasible S_k such that,

$$S_k \geq S_i + d_i - M_{seq}(1 - u_{ik}) \qquad \forall i \in I - \{k\} \tag{2.92}$$

$$S_i \geq S_k + d_k - M_{seq}(1 - u_{ki}) \qquad \forall i \in I - \{k\} \tag{2.93}$$

$$S_k \geq s_k \tag{2.94}$$

Substituting the values of S_i from (2.90) and $M_{seq} = 2M'$ we rewrite the above inequalities to get,

$$S_k \geq S'_i + d_i + d_k + \max_{h \in I} \{s_h\} - M'(1 - u_{ik}) \qquad \forall i \in I - \{k\} \tag{2.95}$$

$$S_k \leq S'_i + \max_{h \in I} \{s_h\} + M'(3 - 2u_{ki} - u_{ik}) \qquad \forall i \in I - \{k\} \tag{2.96}$$

For there to be a feasible S_k, the RHS of (2.95) must be less than or equal to the

RHS of (2.96) i.e.,

$$S_i' + \max_{h \in I} \{s_h\} + M'(3 - 2u_{ki} - u_{ik}) \geq S_i' + d_k + \max_{h \in I} \{s_h\} + d_i \\ - M'(1 - u_{ik}) \; \forall i \in I - \{k\} \tag{2.97}$$

Simplifying the above we get,

$$M'[4 - 2(u_{ki} + u_{ik})] \geq d_k + d_i \qquad \forall i \in I - \{k\} \tag{2.98}$$

As M' is large enough and $u_{ik} + u_{ki} \leq 1$, the above is true. We also need to check that the RHS of (2.96) is greater the RHS of (2.94), i.e.,

$$s_k \leq S_i' + \max_{h \in I} \{s_h\} + M'(3 - 2u_{ki} - u_{ik}) \qquad \forall i \in I - \{k\} \tag{2.99}$$

The above equation is always true as $u_{ik} + u_{ki} \leq 1$. Therefore, there exists feasible S_i $\forall i \in I$ for $I = \{1, 2, \ldots, k\}$ and by principle of mathematical induction, $IARSP^{LP}$ has relatively complete recourse.

Table 2.11: Number of surgeries across specialties

Surgery Specialty	Average Number of Surgeries per day	Range of surgeries per day
Vascular	1.75	0-10
Neuro	5.97	0-18
Plastics	6.01	0-15
ENT	1.53	0-6
Urology	3.23	0-15
Liver	5.15	0-10
Thoracic	1.66	0-12
Cardiac	7.07	1-12
Trauma	1.02	0-7
Pediatric	1.34	0-6
Eye Surgery	1.44	0-5
General	26.34	0-62

Table 2.12: Variance Inflation Factors

Variable	VIF
BOOKEDHRS	3.22
ASA	1.23
AGE	1.15
PATCLASS	1.76
CPTCLUST	4.96
PROVCLUST	3.19
NUMCPT	1.04

Table 2.13: Results of independent variable selection for quantile regression

	Dependent variable:		
	ACTUALHRS		
	(1)	(2)	(3)
BOOKEDHRS	0.206***	0.206***	0.204***
	(0.009)	(0.009)	(0.009)
ASA	0.044***	0.044***	0.036***
	(0.009)	(0.009)	(0.009)
AGE	−0.001***	−0.001***	
	(0.0002)	(0.0003)	
NUMCPT	−0.001*		
	(0.0003)		
CPTCLUST2	−4.966***	−4.964***	−4.954***
	(0.145)	(0.147)	(0.157)
CPTCLUST3	−5.872***	−5.869***	−5.869***
	(0.147)	(0.149)	(0.159)
CPTCLUST4	−6.622***	−6.619***	−6.616***
	(0.148)	(0.150)	(0.160)
CPTCLUST5	−3.726***	−3.725***	−3.720***
	(0.147)	(0.149)	(0.159)
CPTCLUST6	−2.350***	−2.353***	−2.342***
	(0.147)	(0.150)	(0.159)
PROVCLUST2	0.363***	0.367***	0.354***
	(0.066)	(0.067)	(0.067)
PROVCLUST3	0.967***	0.970***	0.962***
	(0.094)	(0.096)	(0.094)
PROVCLUST4	0.484***	0.483***	0.478***
	(0.024)	(0.025)	(0.024)
PROVCLUST5	0.331***	0.330***	0.332***
	(0.015)	(0.016)	(0.016)
PROVCLUST6	0.417***	0.416***	0.400***
	(0.019)	(0.020)	(0.019)
PATCLASS-INPATIENT	0.077**	0.074**	0.078**
	(0.032)	(0.032)	(0.033)
PATCLASS-SAME DAY ADMIT	0.159***	0.159***	0.151***
	(0.031)	(0.031)	(0.032)
Constant	7.245***	7.241***	7.218***
	(0.156)	(0.158)	(0.168)
Observations	19,335	19,335	19,335
AIC	58529	58528	58549
Out-of-sample MSPE	1.312	1.312	1.313
			*p<0.1; **p<0.05; ***p<0.01

Table 2.14: Coefficients of quantile regressions at various quantile levels

	Dependent variable:		
	ACTUALHRS		
	$q = 0.1$	$q = 0.3$	$q = 0.9$
BOOKEDHRS	0.096^{***}	0.151^{***}	0.469^{***}
	(0.017)	(0.009)	(0.020)
ASA	-0.023^{***}	0.011	0.122^{***}
	(0.009)	(0.009)	(0.023)
AGE	-0.002^{***}	-0.002^{***}	-0.0004
	(0.0003)	(0.0002)	(0.001)
CPTCLUST2	-4.337^{***}	-5.027^{***}	-4.563^{***}
	(0.684)	(0.099)	(0.214)
CPTCLUST3	-4.843^{***}	-5.919^{***}	-5.143^{***}
	(0.683)	(0.101)	(0.220)
CPTCLUST4	-5.254^{***}	-6.573^{***}	-6.219^{***}
	(0.684)	(0.103)	(0.225)
CPTCLUST5	-3.420^{***}	-3.783^{***}	-3.433^{***}
	(0.693)	(0.107)	(0.216)
CPTCLUST6	-2.821^{***}	-2.495^{***}	-2.420^{***}
	(0.701)	(0.115)	(0.216)
PROVCLUST2	0.318^{***}	0.297^{***}	0.924^{***}
	(0.064)	(0.079)	(0.123)
PROVCLUST3	0.391^{***}	0.742^{***}	1.812^{***}
	(0.134)	(0.096)	(0.193)
PROVCLUST4	0.228^{***}	0.414^{***}	0.764^{***}
	(0.039)	(0.025)	(0.066)
PROVCLUST5	0.207^{***}	0.295^{***}	0.380^{***}
	(0.026)	(0.018)	(0.047)
PROVCLUST6	0.181^{***}	0.352^{***}	0.538^{***}
	(0.032)	(0.021)	(0.050)
PATCLASS-INPATIENT	0.185^{***}	0.075	0.176
	(0.071)	(0.078)	(0.142)
PATCLASS-SAME DAY ADMIT	0.589^{***}	0.274^{***}	-0.141
	(0.075)	(0.078)	(0.141)
Constant	5.543^{***}	7.130^{***}	7.005^{***}
	(0.691)	(0.136)	(0.282)
Observations	19,335	19,335	19,335
Note:			*p<0.1; **p<0.05; ***p<0.01

Figure 2.7: Histogram of average daily anesthesia hours

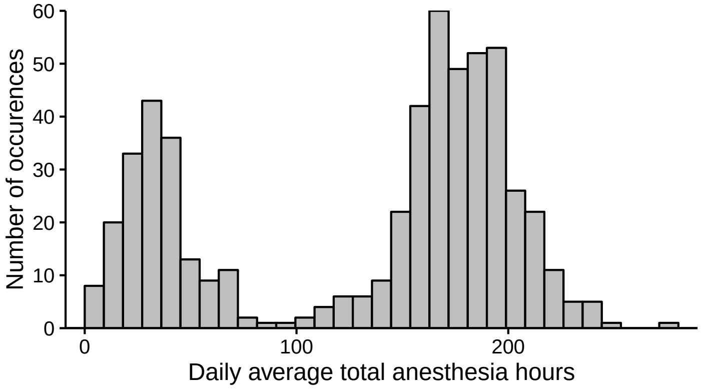


Figure 2.8: Histogram of average daily anesthesia hours controlling for number and mix of surgeries

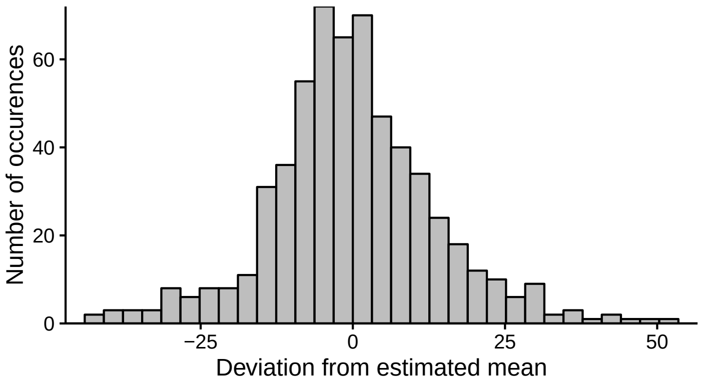


Figure 2.9: Number of rooms with assigned anesthesiologists by hour of day

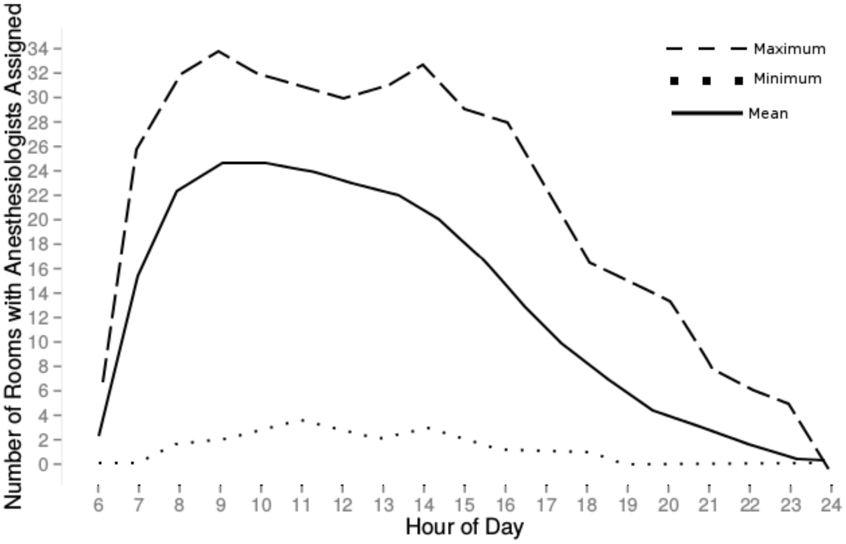


Figure 2.10: Number of CPT code clusters and % explained variance

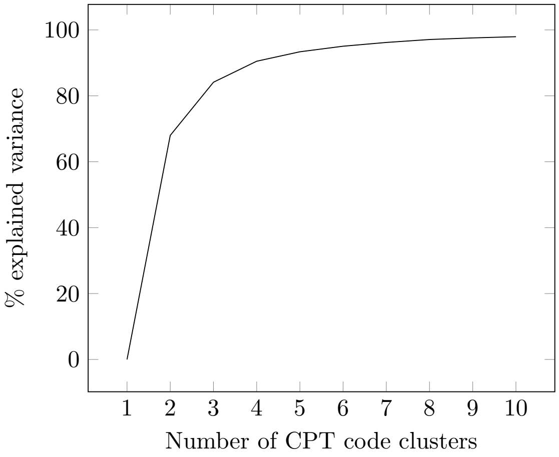


Figure 2.11: Number of surgeon clusters and % explained variance

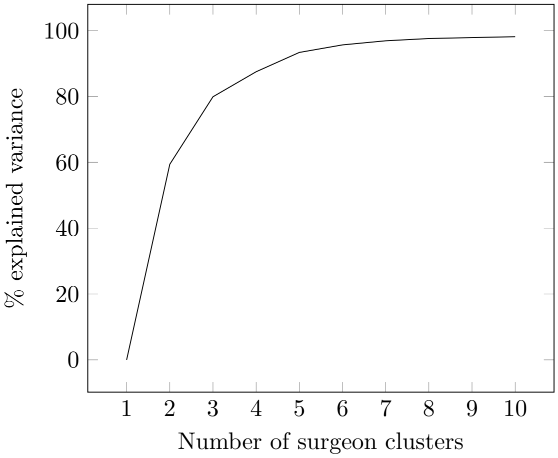
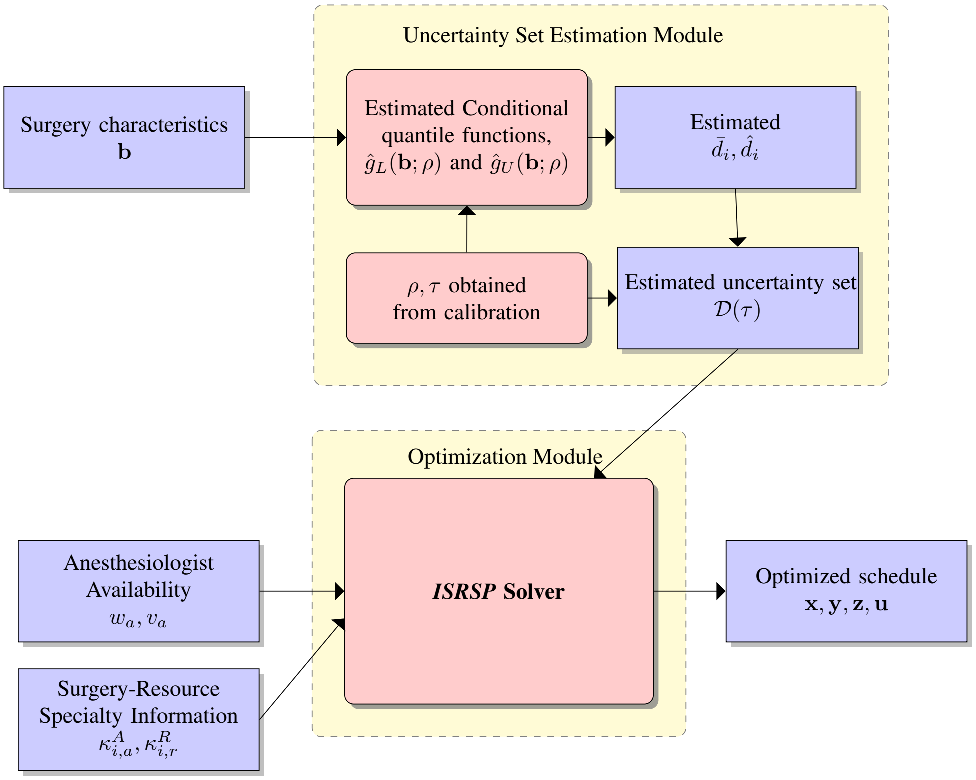


Figure 2.12: Software schematic of decision support system



Uncertainty Set Estimation Module
Surgery characteristics $\mathbf{b}$
Estimated Conditional quantile functions, $\hat{g}_L(\mathbf{b};\rho)$ and $\hat{g}_U(\mathbf{b};\rho)$
Estimated $\bar{d}_i, \hat{d}_i$
ρ, τ obtained from calibration
Estimated uncertainty set $\mathcal{D}(\tau)$
Optimization Module
Anesthesiologist Availability w_a, v_a
ISRSP Solver
Optimized schedule $\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{u}$
Surgery-Resource Specialty Information $\kappa^A_{i,a}, \kappa^R_{i,r}$

CHAPTER 3

Planning for HIV Screening, Testing and Care at the Veterans Health Administration

3.1 Introduction

Veterans Health Administration (VHA), one of the components of the Veterans Administration, is the largest integrated healthcare provider in the United States of America (USA). The VHA is funded by the federal government and serves the medical and social support needs of over 8 million active duty and honorably discharged veterans over their entire lifetime. The VHA provides these services through 128 stations. For the purpose of this paper, we shall focus on the Greater Los Angeles (GLA) station, as the unit of analysis because of our close working relationship with its key decision makers.

The VHA is the largest provider of HIV care in the USA. As of 2011, the VHA reported over 25,271 HIV infected patients, an increase of 3.7% from 2007. The VHA is also a leader in quality of care provided to HIV infected patients with high adherence to the Department of Health and Human Services clinical guidelines across all regions. An important aspect of HIV care is early diagnosis and treatment which is known to lower cost and improve patient outcomes (Palella et al., 2003). In addition, this reduces the incidence of secondary complications which are very costly to treat if HIV itself is not treated in a timely manner

(Schackman et al. 2006). Prior studies at the VHA (Nayak et al., 2012) show that a major factor impeding the early diagnosis and treatment of HIV is the policy of risk-based screening. Under this policy, patients are tested for HIV only if they display certain risk factors such as injection drug use, or if they present symptoms of opportunistic infections. Owens et al. (2007) found that only 36% of at risk patients had ever been tested for HIV. The main operational barriers cited for insufficient coverage of screening and late diagnosis of HIV infection were constraints on provider time and insufficient capacity of trained counselors (Goetz et al., 2008a).

An alternative policy recommended by the Centers for Disease Control (CDC) is to implement routine HIV screening, in which a patient visiting the healthcare facility would be offered an HIV test irrespective of risk factors or symptoms. Several recent studies in the public health literature have found that such routine HIV screening is cost-effective [1] compared to risk based testing even in settings with very low prevalence of HIV (Paltiel et al., 2005). In 2009, the VHA proposed to implement the routine screening policy across its stations [2]. Consequently, the management at the GLA station wanted to understand if such a policy would be feasible given their capacity and budgetary constraints, and if necessary, was willing to consider alternate policies to improve upon their current risk based screening policy. In response, we developed an optimization model to achieve these goals at the GLA station. Consistent with the mission of the VHA of providing high quality care over the lifetime of veterans, the objective of this model is to maximize the total QALYs of all the patients at this station. To achieve this objective, this model determines the optimal fraction of patients to

[1] A policy or intervention is said to be cost effective if the Quality Adjusted Life Years (QALYs) gained due to that intervention costs less than $109,000 to $297,000 per QALY gained. (http://www.cdc.gov/hiv/prevention/ongoing/costeffectiveness/). The term QALYs is commonly used in the health economics and health policy literature to assess the value of a medical intervention in terms of the number of years at a particular quality level added due to the intervention (Dolan et al. 2005).

[2] http://www.va.gov/vhapublications/ViewPublication.asp?pub_ID=2056

be screened (i.e., offered the test) and also determines the optimum staffing levels at different parts or locations of the station. This model explicitly captures patient flow and the associated disease progression through system dynamics constraints. In addition, it also incorporates budget and capacity constraints.

We first used this model to evaluate the current risk based screening policy and the proposed routine screening policy at the GLA station. We found that the cost-effective routine screening policy was not feasible in the current budgetary environment at this station. Therefore, we developed four other policies within the framework of our model that improved upon the current risk based screening policy. An extensive computational analysis provided a benchmark value for each policy and provided guidance in terms of the fraction of patients to be screened in every period as well as the number of healthcare workers that need to be staffed at each part of the system in order to implement a policy. Thus, unlike conventional cost effective analysis, our approach provided a feasible plan that can be implemented.

Optimization based models have been used to evaluate prevention and treatment policies for HIV at different decision making levels (Rauner and Brandeau, 2001). Population level studies evaluate the cost effectiveness of policy interventions (Zaric and Brandeau, 2001; Long et al., 2010), while studies at an individual patient level optimize clinical decision making to maximize patient welfare (Roberts et al., 2010; Shechter et al., 2008). Healthcare systems face the problem of integrating cost effective policies with clinical decisions subject to organizational and budgetary constraints. Blount et al. (1997), Zaric et al. (2000) and Brandeau et al. (2003) evaluate general formulations of this problem with budget constraints to decide optimal intervention for prevention of infectious diseases. Their approximations lead to formulations that can be solved by linear programming and convex optimization techniques. More recently, Kucukyazici et al. (2011) and Deo et al. (2013) combine clinical models of disease progression for chronic diseases with operational models of the health system. However, none of these papers consider different parts of the healthcare system with capacity constraints and do not jointly optimize screening and staffing decisions, which are the key features of the decision problem faced by the VHA.

Our paper makes the following contributions. First, it models a very relevant but complex problem at the interface of operations management and public health. It then develops methods for the efficient computation of bounds and managerially relevant solutions for this problem. Second, to the best of our knowledge, this is the first planning model which determines the fraction of patients that need to be screened along with the staffing requirements at screening, testing and care, while including disease progression and flow of patients in different health states across various parts of a constrained healthcare system. Third, we explicitly consider capacity and budget constraints and illustrate their impact on screening and staff allocation decisions. Fourth, we apply the model to data collected from the GLA station to analyze various policies. Our computational analysis shows that GLA station can achieve substantial increase (20% to 300%) in the QALYs gained by using these policies and our model provides guidance for its effective implementation. Fifth, the insights from our model have influenced planning decisions at this station. In addition, two policies have been used at the GLA station and our analysis provides the basis to extend and enhance these policies.

The remainder of the paper is organized as follows. In Section 2, we describe the healthcare system, patient health states, disease progression and system dynamics. These form the basis of our optimization model, which is formulated in Section 3. We also discuss structural properties, construct an upper bound and develop four policies that serve as lower bounds for this model. In Section 4, we describe various primary and secondary sources of data used in the model. Section 5 analyzes several policies for HIV screening, testing and care that can be evaluated within the framework of our model. Section 6 describes the application and qualitative impact of this work.

3.2 Problem Description

The GLA station is one of the largest and the most complex stations in the VHA consisting of 3 ambulatory care centers, a tertiary care facility and 10 community based clinics. The GLA serves veterans residing in Los Angeles, Kern, Santa Barbara, Ventura and San Louis Obispo counties. The GLA station management recommended that we conduct a station level analysis because it was difficult to estimate the budget for individual facilities within the station. Further, the management felt that such an analysis could lead to effective staff reallocation because there was considerable flexibility in adjusting the staffing levels across facilities within a station. From a managerial perspective, these aspects were considered more important than any potential downside due to loss of granularity in terms of patient flow and staffing.

As discussed before, the primary benefit of routine screening is early diagnosis of HIV positive patients and their connection to care before they become symptomatic. This benefit arises from the fact that the healthcare cost of asymptomatic HIV patients (including HIV treatment and other hospitalization) is much lower and their quality of life is much better than that of symptomatic HIV patients (Kaplan et al. 2009). In order to capture this effect, we constructed a compartmental model of patients with each compartment corresponding to a combination of the health state of the patients and part of the healthcare system to which they belong. Below, we describe the healthcare system, patient health states, disease progression, and system dynamics.

3.2.1 Healthcare System

Based on our discussions with the station management, we divided the healthcare system at the station into three distinct parts: 1) primary care (facilities such as outpatient clinics and hospitals where patients are screened or are offered an HIV test and blood samples are collected if they agree to be tested), 2) laboratory (a central location where samples collected during screening are tested), and 3) infectious disease specialty care (where HIV positive patients are referred for monitoring or treatment). Primary and specialty care could be staffed by up to three worker types: physicians, nurses and counselors, while the laboratory is only staffed by the laboratory technician. Staffing levels are fixed during the budget horizon of one year to provide certainty and foster a stable work environment for all their staff. To provide a precise definition of the healthcare system, let $\tau \in [T]$ denote the budget periods, each corresponding to a year and let $t \in \mathcal{M}_\tau = \{1 + 12(\tau - 1), \ldots, 12\tau\}$ index the set of discrete time periods corresponding to a month within the budget period. Further, let $k \in \mathcal{W} = \{phys, nurse, couns, lab\}$ index the set of worker types, and $\ell \in \mathcal{L} = \{P, L, S\}$ index the set of parts or locations where P denotes primary care facility, L denotes laboratory and S denotes infectious diseases specialty care. Each location ℓ is staffed by $n_{k,\ell}$ healthcare workers of type k, each of whom earns a wage w_k in each period and spends a total of $y_{k,\ell}$ time units on average with the patient. Since the healthcare workers Since the healthcare workers have other tasks associated with other diseases and conditions, we assume that the total time available with the resource of type k in location ℓ for the HIV routine screening program is limited and denoted by $A_{k,\ell}$.

3.2.2 Patient Health States

Following earlier work in the modeling of disease progression in HIV patients, Freedberg et al. (1998) (Mauskopf et al., 2005), we use different ranges of CD4 cell count[3], and the presence or absence of Opportunistic Infections (OI) to define a set of health states of HIV infected patients. In addition, we include uninfected and dead as two additional health states. Table 3.1 below provides the definition of the resulting 14 health states based on CD4 count range and their associated states of OI. These states are indexed by i and j in the model.

Table 3.1: CD4 Health States

Health state index (i, j)	CD4 count range (cells/mm^3) *without* opportunistic infections	Health state index (i, j)	CD4 count range (cells/mm^3) *without* opportunistic infections
0	Uninfected	7	500+
1	500+	8	350-499
2	350-499	9	200-349
3	200-349	10	100-199
4	100-199	11	50-99
5	50-99	12	0-49
6	0-49	13	Death

[3] CD4$^+$T helper cells are white blood cells essential to the human immune system and are usually expressed as number of cells per milliliter. Patients infected with HIV show reduced number of CD4 cells and a lower number of CD4 indicates a greater progression of the infection.

In addition, the VHA identifies incoming patients as either high risk or low risk depending on their observable characteristics such as previous Hepatitis B or C infection, injection drug use, or homelessness. These risk categories are indexed by $r \in \mathcal{R} = \{1, 2\}$, where $r = 1$ signifies patients of higher risk of infection of HIV and $r = 2$ signifies those with a lower risk of infection. At the GLA station, 25% of the patients were classified as high risk, and the remaining 75% were classified as low risk (Goetz et al., 2013).

3.2.3 Disease Progression

In single patient models, the transition between health states is typically modeled as a discrete time Markov chain in which the probability of transitioning from state to state is conditionally independent of the history of earlier transitions. However, this approach is analytically intractable for a multi-period aggregate or population-level model like ours, which also considers multiple parts of the health care system while optimizing screening and staff allocation decisions. Hence, we approximate the disease progression model by using deterministic transition rates in which we assume that a fixed fraction of the number of patients move from one health state to the other in each period[A similar approach is used in mathematical epidemiology to model the spread of infectious diseases in the population (Anderson et al., 1992).] This deterministic approximation of transition rates is reasonable here since the unit of our analysis is the GLA station and the population of patients in each state is relatively large. We use $\theta^{i,j}_{r,\omega}$ to denote the fraction of patients in health state i that move to health state j in one month. This fraction depends on the patient risk category r , and the treatment status $\omega \in \Omega = \{treat, untreat\}$, where *treat* refers to undergoing antiretroviral treatment and *untreat* represents not undergoing treatment respectively.

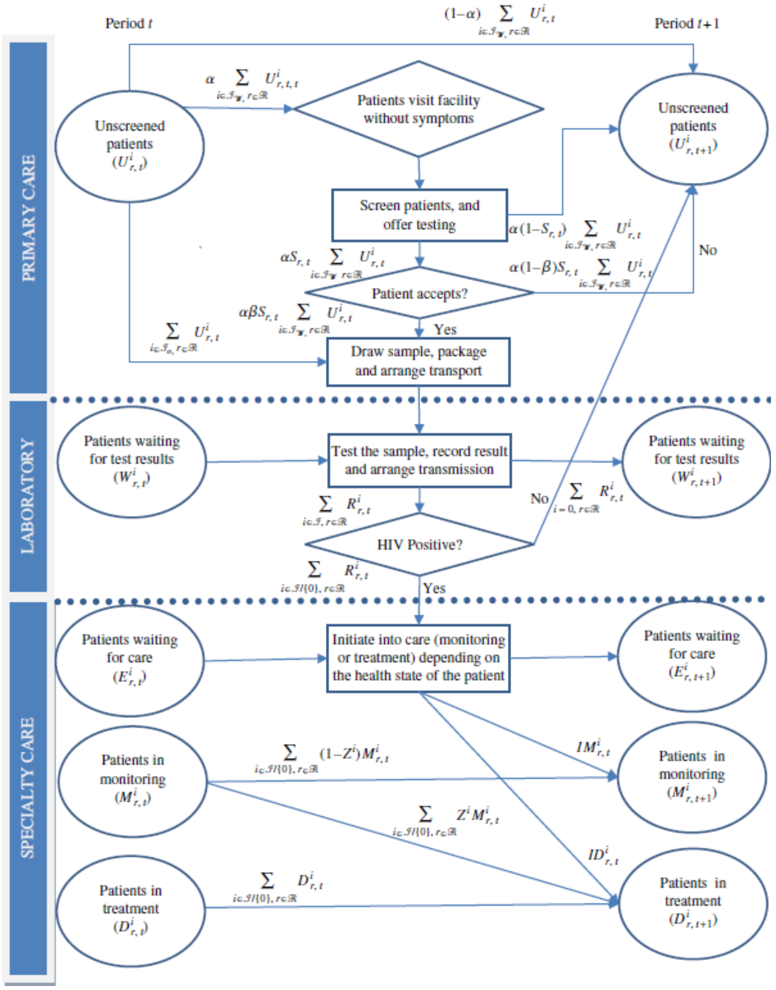
Four processes govern the transition across health states: 1) HIV infection,

2) HIV infection progression (treated and untreated), 3) Opportunistic infection (OI), and 4) OI recovery. We used clinical data to estimate the transition rates associated with each of these processes separately. For certain transitions that require more than one process simultaneously, we assumed that the rate of one process does not depend on the other. Details on the calculations of the transition rates are provided in the Appendix.

3.2.4 System Dynamics

In this section, we describe the system dynamics obtained by combining disease progression with patient flows to represent how patients move across different health states as well as various parts of the health care system over time. In particular, we track the number of patients in each risk category r, each health state i, at each location ℓ, in each time period t. Figure 3.1 shows the flow of patients through various parts of the health care system.

Figure 3.1: Flow of patients through different parts of the health care system in the Greater Los Angeles


Period t
$(1-\alpha) \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} U^i_{r,t}$
Period $t+1$
PRIMARY CARE
$\alpha \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} U^i_{r,t,t}$
Patients visit facility without symptoms
Unscreened patients $(U^i_{r,t})$
Unscreened patients $(U^i_{r,t+1})$
Screen patients, and offer testing
$\alpha(1-S_{r,t}) \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} U^i_{r,t}$
No
$\alpha S_{r,t} \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} U^i_{r,t}$
$\alpha(1-\beta) S_{r,t} \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} U^i_{r,t}$
Patient accepts?
$\alpha\beta S_{r,t} \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} U^i_{r,t}$
$\sum_{i \in \mathcal{I}_o, r \in \mathcal{R}} U^i_{r,t}$
Yes
Draw sample, package and arrange transport
LABORATORY
Patients waiting for test results $(W^i_{r,t})$
Test the sample, record result and arrange transmission
Patients waiting for test results $(W^i_{r,t+1})$
$\sum_{i \in \mathcal{I}, r \in \mathcal{R}} R^i_{r,t}$
No
$\sum_{i=0, r \in \mathcal{R}} R^i_{r,t}$
HIV Positive?
$\sum_{i \in \mathcal{I}/\{0\}, r \in \mathcal{R}} R^i_{r,t}$
Yes
SPECIALTY CARE
Patients waiting for care $(E^i_{r,t})$
Initiate into care (monitoring or treatment) depending on the health state of the patient
Patients waiting for care $(E^i_{r,t+1})$
$IM^i_{r,t}$
Patients in monitoring $(M^i_{r,t})$
$\sum_{i \in \mathcal{I}/\{0\}, r \in \mathcal{R}} (1-Z^i) M^i_{r,t}$
Patients in monitoring $(M^i_{r,t+1})$
$\sum_{i \in \mathcal{I}/\{0\}, r \in \mathcal{R}} Z^i M^i_{r,t}$
$ID^i_{r,t}$
Patients in treatment $(D^i_{r,t})$
$\sum_{i \in \mathcal{I}/\{0\}, r \in \mathcal{R}} D^i_{r,t}$
Patients in treatment $(D^i_{r,t+1})$

Primary Care – Screening. The process starts with patients who are unaware of their HIV status, whom we call unscreened patients. Let $U^i_{r,t}$ denote the total number of unscreened patients in risk category, health state i and at time period t. All patients with an opportunistic infection ($i \in \mathcal{I}_o = \{7, 8, \ldots, 13\}$) are immediately offered the HIV test and their acceptance rate is 100%. A fraction α of the remaining asymptomatic patients who do not have OI ($i \in \mathcal{I}_w = \{0, 1, 2, \ldots, 6\}$)

visit a primary care facility in period t for other conditions. Let $S_{r,t}$ represent the fraction of patients of risk category r in period t that are screened or offered the HIV test. A fraction β of these patients accepts the test. The number of unscreened patients in the next time period $U^i_{r,t+1}$ is given by,

$$U^i_{r,t+1} = \left(\sum_{j \in \mathcal{I}_w} \theta^{i,j}_{r,untreat}(1 - \alpha\beta S_{r,t}) U^j_{r,t} \right) + N^i_{r,t+1} + R^0_{r,t}\theta^{0,i}_{r,untreat} \quad \forall r, i, t \tag{3.1}$$

The first term $\left(\sum_{j \in \mathcal{I}_w} \theta^{i,j}_{r,untreat}(1 - \alpha\beta S_{r,t}) U^j_{r,t}\right)$ of this equation is derived by summing three types of patient flows shown in Figure 3.1. (a) the asymptomatic patients who do not visit the clinic; (b) those who visit and do not get screened; and (c) those who visit, get selected for a test, and refuse to be tested. This sum is appropriately weighted by the rates of transition from state j to state i as determined by the disease progression model. The second term ($N^i_{r,t+1}$) is the number of new patients in health state i and risk category r who enter in period $(t+1)$. The third term ($R^0_{r,t}\theta^{0,i}_{r,untreat}$) is the number of uninfected patients who receive a negative HIV test at the beginning of period t and join the pool of unscreened population in the next period.

Laboratory–Testing The blood samples collected from patients who accept the offered test are then sent to the lab where the actual test is conducted and the results are communicated back to the patient. Here, we allow for a lag between the collection of the sample and return of the results due to congestion at the lab. Let $W^i_{r,t+1}$ represent the number of patients in health state i and risk category r who are waiting to receive their results at the beginning of the period $t+1$ in the laboratory. This is given by

$$\begin{aligned} W^i_{r,t+1} = & \sum_{j \in \mathcal{I}} W^j_{r,t}\theta^{j,i}_{r,untreat} + \sum_{j \in \mathcal{I}_w} \alpha\beta S_{r,t} U^j_{r,t}\theta^{j,i}_{r,untreat} + \sum_{j \in \mathcal{I}_o} U^j_{r,t}\theta^{j,i}_{r,untreat} \\ & - \sum_{j \in \mathcal{I}} R^j_{r,t}\theta^{j,i}_{r,untreat} \qquad \forall r, i, t \end{aligned} \tag{3.2}$$

$W^i_{r,t+1}$ consists of four terms. The first term, $(\sum_{j\in\mathcal{I}} W^j_{r,t}\theta^{j,i}_{r,untreat})$ represents the number of patients waiting at the beginning of period t who have undergone disease progression, where $\mathcal{I} = \mathcal{I}_o \cup \mathcal{I}_w$. The second term $(\sum_{j\in\mathcal{I}_w} \alpha\beta S_{r,t} U^j_{r,t}\theta^{j,i}_{r,untreat})$, represents the number of asymptomatic patients who accept the test offer at the beginning of period t. The third term, $(\sum_{j\in\mathcal{I}_o} U^j_{r,t}\theta^{j,i}_{r,untreat})$ represents the number of symptomatic patients who directly proceed to testing. The fourth term $(\sum_{j\in\mathcal{I}} R^j_{r,t}\theta^{j,i}_{r,untreat})$ represents the patients who receive their results and who either exit the system because their tests are negative (i.e., $j = 0$) or who are now transferred to care (i.e., $j \neq 0$). As before, multiplication by $\theta^{j,i}_{r,untreat}$ in each term represents disease progression in one period.

Specialty Care–Monitoring and Treatment Patients who receive positive test results are connected to infectious diseases specialty care for monitoring and treatment. Again, we allow for a lag between the receipt of results and being connected to care. Let $I^i_{r,t}$ denote the number of patients of risk category r and health state i who are initiated into care. Of these, depending on the stage of their disease progression, $IM^i_{r,t}$ are initiated under monitoring and $ID^i_{r,t}$ are immediately initiated into treatment. Let $E^i_{r,t+1}$ denote the number of patients at the beginning of the period $t+1$ who are waiting to be enrolled in care. This is given by

$$\begin{aligned} E_{r,t+1} = & \sum_{j\in\mathcal{I}-\{0\}} R^j_{r,t}\theta^{i,j}_{r,untreat} + \sum_{j\in\mathcal{I}-\{0\}} E^j_{r,t}\theta^{i,j}_{r,untreat} \\ & - \sum_{j\in\mathcal{I}-\{0\}} IM^j_{r,t}\theta^{i,j}_{r,untreat} - \sum_{j\in\mathcal{I}-\{0\}} ID^j_{r,t}\theta^{i,j}_{r,treat} \qquad \forall r, i, t \end{aligned} \tag{3.3}$$

The first term $(\sum_{j\in\mathcal{I}-\{0\}} R^j_{r,t}\theta^{i,j}_{r,untreat})$ is the number of patients who received positive HIV test results at the beginning of period t. The second term, $\sum_{j\in\mathcal{I}-\{0\}} E^j_{r,t}\theta^{i,j}_{r,untreat}$ is the number of patients who were waiting to be enrolled into care at the beginning of period t. The third and fourth terms $(\sum_{j\in\mathcal{I}-\{0\}} IM^j_{r,t}\theta^{i,j}_{r,untreat}, \sum_{j\in\mathcal{I}-\{0\}} ID^j_{r,t}\theta^{i,j}_{r,treat})$ are the number of people who were enrolled at period t into monitoring and treatment, respectively. Patients who are enrolled into treatment now undergo disease progression under the parameter $\theta^{i,j}_{r,treat}$ instead of $\theta^{i,j}_{r,untreat}$. The decision to initiate patients under monitoring or under treatment depends on the health state of the patient and current clinical guidelines described in §3.4.2. We use a binary indicator parameter z^i to capture the clinical decision whether all patients at health state i are initiated under treatment ($z^i = 1$) or monitoring ($z^i = 0$). Then the number of patients who are initiated into treatment and monitoring at time period t is given by the following equations:

$$ID^i_{r,t} = I^i_{r,t} z^i \qquad \forall r, i, t \tag{3.4}$$

$$IM^i_{r,t} = I^i_{r,t}(1 - z^i) \qquad \forall r, i, t \tag{3.5}$$

Next, consider $M^i_{r,t+1}$, the number of patients of risk category r under monitoring in state i at the beginning of time period $t + 1$, this is given by,

$$\begin{aligned} M^i_{r,t+1} = &\sum_{j \in \mathcal{I} - \{0\}} M^j_{r,t} \theta^{j,i}_{r,untreat} - \sum_{j \in \mathcal{I} - \{0\}} M^j_{r,t} z^j \theta^{j,i}_{r,treat} \\ &+ \sum_{j \in \mathcal{I} - \{0\}} IM^j_{r,t} \theta^{j,i}_{r,untreat} \qquad \forall r, i, t \end{aligned} \tag{3.6}$$

The first term in Equation 3.6 represents the number of patients in health state i who remain under monitoring at the beginning of period t, the second term represents those who enter treatment from monitoring, and the third term represents the newly diagnosed patients who enter care under monitoring.

Finally, let $D^i_{r,t+1}$ represent the number of patients under treatment in state i at the beginning of period $t + 1$. This is given by

$$\begin{aligned} D^i_{r,t+1} = &\sum_{j \in \mathcal{I} - \{0\}} D^j_{r,t} \theta^{j,i}_{r,treat} + \sum_{j \in \mathcal{I} - \{0\}} M^j_{r,t} z^j \theta^{j,i}_{r,treat} \\ &+ \sum_{j \in \mathcal{I} - \{0\}} ID^j_{r,t} \theta^{j,i}_{r,treat} \qquad \forall r, i, t \end{aligned} \tag{3.7}$$

The first term in Equation 3.7 represents the number of patients under treatment in period t in a particular health state, the second term denotes the number of patients who enter treatment from the pool of monitored patients, and the third term is the number of newly diagnosed patients who enter treatment. In formulating the system dynamics we have made the following simplifying assumptions. First, once patients enter the system and are tested, they can exit the system only if they are uninfected or if they die. Second, all primary care locations fully comply with the screening policy. Third, the treatment protocol is well defined and is followed by all physicians at the infectious diseases specialty care. These assumptions were validated by prior internal studies at the GLA station. Given the health care system, patient health states, disease progression, and system dynamics the overall objective of the GLA station is to maximize the aggregated QALYs across all patients in the system. This can be done by appropriately choosing the screening fraction and consequently the number of patients to be screened, tested, and cared for in every period and by determining the staffing level at each part of the healthcare system to execute this choice. While doing this, the station faces organizational constraints relating to capacity and budget availability. We next develop an optimization model for this decision problem.

3.3 Model

In this section, we start by describing the objective function and the organizational constraints related to budget and capacity. These together with the previously described system dynamics form a discrete time planning model. We characterize key properties of this model and use them to develop an upper bound that can be employed to evaluate the quality of any given solution. Finally, we develop managerially relevant heuristics or policies to solve this model. Table 3.2 summarizes all notations that are used in the model, including those that have already been introduced in the previous section.

Table 3.2: Notation for Planning for HIV Screening, Testing and Care Problem

Symbol	Description
	Indices
$\tau \in [T] = \{1, 2, \ldots, T\}$	Number of years
$t \in \mathcal{M}_\tau = \{1 + 12(\tau - 1), \ldots, \tau\}$	Number of months
$k \in \mathcal{W} = \{phys, nurse, couns, lab\}$	Resource type
$\ell \in \mathcal{L} = \{P, L, S\}$	Location within healthcare system, P:Primary care facility, L:Laboratory, S: Infectious diseases sub-specialty
$i, j \in \mathcal{I}_w = \{0, 1, \ldots, 6\}$	Health states corresponding to patients without OI
$i, j \in \mathcal{I}_o = \{7, 1, \ldots, 13\}$	Health states corresponding to people with OI
$i, j \in \mathcal{I}_w \cup \mathcal{I}_o = \mathcal{I}$	Health states of all patients
$\omega \in \Omega = \{treat, untreat\}$	Treatment status
$r \in \{1, 2\}$	Risk category
$X \in \mathcal{X} = \{U, W, E, M, D\}$	System state: U : Unscreened, W : Waiting for results, E: Waiting to be enrolled into monitoring or treatment, M: Monitoring, D: Treatment
	Parameters related to patient flow
$\hat{p}_r^i$	Fraction of patients in risk category r of health state i in the new patient population

Continued on next page

Table 3.2 – continued from previous page

Symbol	Description
α	Fraction of asymptomatic patients who visit healthcare facility
β	Fraction of patients who accept screening
$\theta^{i,j}_{r,\omega}$	Fraction of patients in risk category r and under treatment status moving from health state i to health state j in one month
q^i	Quality of life score for patients in health state i
N_i	Number of new patients entering the system in period t
z^i	A binary parameter indicating whether patient of health state i is initiated under monitoring ($z^i = 0$) or treatment ($z^i = 1$)
	Parameters related resource utilization
$y_{k,\ell}$	Time required per patient of healthcare worker of type k at location l
$A_{k,\ell}$	Total time available for HIV screening program of healthcare worker of type k at location l
w_k	Per period wages of healthcare worker of type k
CS^i	Cost of screening per patient

Continued on next page

Table 3.2 – continued from previous page

Symbol	Description
C_X^i	Cost per patient in system state X
$B(\tau)$	Total annual budget available for HIV related activities in year τ
	State variables
$U_{r,t}^i$	Number of unscreened patients of risk category r in health state i at the beginning of period t
$W_{r,t}^i$	Number of patients of risk category r in health state i waiting for their results at the beginning of period t
$R_{r,t}^i$	Number of patients of risk category r in health state i who receive their results in period t
$E_{r,t}^i$	Number of patients of risk category r in health state i waiting to be enrolled at the beginning of period t
$M_{r,t}^i$	Number of patients of risk category r in health state i who are under monitoring at the beginning of period t
$D_{r,t}^i$	Number of patients of risk category r in health state i who are under treatment at the beginning of period t
$ID_{r,t}^i$	Number of patients of risk category r in health state i who are initiated under treatment in period t

Continued on next page

Table 3.2 – continued from previous page

Symbol	Description
$IM^i_{r,t}$	Number of patients of risk category r in health state i who are initiated under monitoring in period t
$I^i_{r,t}$	Number of patients of risk category r who are initiated under care (monitoring and treatment) in period t
	Decision variables
$S_{r,t}$	Fraction of asymptomatic patients of risk category r visiting a primary care facility in period t who are screened or offered the HIV test
$n_{k,l}$	Number of healthcare workers of type k to be staffed at location ℓ

3.3.1 Objective Function

In accordance with the existing literature on economic evaluation of health interventions and programs (Dolan et al., 2005), we choose the objective function of maximizing the total QALYs gained for the entire patient population over the problem horizon. Note that using this measure ensures that aggregate survival as well as quality of life of patients is considered. Although QALYs is not an operational metric that is used regularly for planning and scheduling decisions within the VHA, this seemed a reasonable objective because it is consistent with the mission of the VHA.

Calculating QALYs involves first associating each health state i with a quality of life (QOL) utility q_i and then multiplying the QOL utility of each health state with the corresponding number of patients in that state. These are calculated by using Equations (3.1) through (3.7) developed in §3.2.4. The QOL utility is a measure of health related utility of patients and ranges between 0 and 1, where 0 corresponds to death and 1 corresponds to perfect health. Finally, the total QALYs are calculated over the entire period of analysis. Using this approach, the objective function can be represented by

$$\sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i\left(U^i_{r,t}+W^i_{r,t}+E^i_{r,t}+M^i_{r,t}+D^i_{r,t}\right)$$

3.3.2 Organizational Constraints

We consider two main sources of organizational constraints in our model. The first is concerned with total annual HIV related budget at the level of a station, and the second defines service level constraints in various parts of the healthcare system within the station.

Budget Constraint The budget at the GLA station consists of three components: the screening cost, healthcare costs associated with a patient in a particular system state, and the cost of wages. This is represented by the following set of inequalities:

$$\begin{aligned}&\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i\alpha\beta S_{r,t}U^i_{r,t}+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^iUr,t^i\\ &\qquad+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau,X\in\mathcal{X}} C^i_XX^i_{r,t}+\sum_{\ell\in\mathcal{L},k\in\mathcal{W},t\in\mathcal{M}\mathcal{T}} n_{k,l}w_k\le B(\tau)\qquad\forall\tau\qquad(3.8)\end{aligned}$$

The first two terms in Equation (3.8) correspond to the screening costs. This is obtained by multiplying the cost of screening per patient in health state i (CS^i) with $\sum_{i\in\mathcal{I}_w}\alpha\beta U^i_{r,t}$, representing the asymptomatic patients who accepted the offered HIV test and with $\sum_{i \in \mathcal{I}_o} U^i_{r,t}$ denoting the number of symptomatic HIV patients who were transferred straight to testing. Both these terms are aggregated across all risk categories and time periods up to one year. The third term represents the cost of providing healthcare services to patients in different system states. This cost is composed of several components that depend on the system state of the patient. For example, if a patient is in treatment, the cost components would be pharmacy, testing, inpatient, outpatient, and overhead costs. Further, the magnitude of this component will also depend on the health state of the patients. For instance, more critically ill patients with lower CD4 counts would typically incur higher pharmacy costs. We combine all such cost components into one parameter, C^i_X representing the cost of having one patient in health state i at system state X. Here, $X \in \mathcal{X} = \{U, W, E, M, D\} =$ {Unscreened, Waiting for results, Waiting to be enrolled, Monitoring, Treatment}. The fourth term in the equation above is the labor cost, which is the salary by resource type k multiplied by the staffing level of that resource type at a particular location ℓ.

Service Level Constraints In addition to the budget constraint, the GLA station would also like to ensure timely service of patients and avoid long delays. We model this requirement using a constraint $P\{W_\ell \leq \tau_\ell\} \geq \alpha_\ell \;\; \forall \ell \in \mathcal{L}$. Where W_ℓ is the random waiting time at location ℓ. This can be interpreted as the probability that the waiting time is less than a specified quantity τ_ℓ and must be greater than a certain threshold α_ℓ. Here, the tuple (τ_ℓ, α_ℓ) was specified at each location based on the organizational goals at the VHA. We use a $M/M/1$ queuing model to approximate $P\{W_\ell \leq \tau_\ell\} = 1 - \exp(-\tau_\ell(\mu_\ell - \lambda_\ell)) \quad \forall \ell \in \mathcal{L}$ (Kleinrock, 1975). Here, λ_ℓ denotes the arrival rate at location ℓ, whereas μ_ℓ denotes the service rate at location ℓ. Using the natural logarithm operator, this can be reformulated as,

$$\lambda_\ell \leq \mu_\ell + \frac{1}{\tau_\ell} \ln(1 - \alpha_\ell) \qquad \forall \ell \tag{3.9}$$

Since the second term on the right-hand side of constraint (3.9) is negative, this constraint is tighter than the traditional capacity feasibility condition $\lambda_\ell \leq \mu_\ell$, which does not impose any requirements on waiting times. Note that reducing quantity τ_ℓ or increasing threshold α_ℓ reduces the effective capacity $\bar{\mu}_\ell - \mu_\ell + (1/\tau_\ell)\ln(1 - \alpha_\ell)$ and further tightens this constraint. To operationalize (3.9) we need to compute (λ_ℓ, μ_ℓ) $\forall \ell \in \mathcal{L}$. The capacity of resource k at location ℓ is given by $n_{k,\ell}A_{k,\ell}/y_{k,\ell}$ patients. Therefore, we approximate the service rate at location ℓ as the minimum or bottleneck capacity across all the resource or worker types available at that location given by $\mu_\ell = \min_k \{n_{k,\ell}A_{k,\ell}/y_{k,\ell}\}$. Below we use the system dynamics developed in §3.2.4 to calculate λ_ℓ are derive the service level constraints for each location.

Primary Care: $(\ell = P)$. Observe from Figure 3.1 that the number of patients to be screened in period t is given by t is given by $\sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t}$. Therefore, $\lambda_P = \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t}$. Therefore, $\lambda_P = \sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i \in \mathcal{I}_o - \{13\}, r \in \mathcal{R}} U^i_{r,t}$ and $\mu_P = \min_k \{n_{k,P}A_{k,P}/y_{k,P}\}$ Substituting these in inequality (3.9), we get the service level constraint for screening as,

$$\sum_{i \in \mathcal{I}_w, r \in \mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i \in \mathcal{I}_o - \{13\}, r \in \mathcal{R}} U^i_{r,t} \leq \min_k \{n_{k,P}A_{k,P}/y_{k,P}\} + \frac{1}{\tau_P} \ln(1 - \alpha_P) \quad \forall t \tag{3.10}$$

Laboratory: $(\ell \in L)$. Figure 3.1 shows that the number of patients who receive their results is $\sum_{i \in \mathcal{I} - \{13\}, r \in \mathcal{R}} R^i_{r,t}$, which is also the input rate, under the assumption of stability. Therefore, $\lambda_L = \sum_{i \in \mathcal{I} - \{13\}, r \in \mathcal{R}} R^i_{r,t}$ and $\mu_L = \min_k \{n_{k,L}A_{k,L}/y_{k,L}\}$. Substituting these in inequality (3.9), we get the service level constraint for L as,

$$\sum_{i \in \mathcal{I} - \{13\}, r \in \mathcal{R}} R^i_{r,t} \leq \min_k \{n_{k,L}A_{k,L}/y_{k,L}\} + \frac{1}{\tau_L} \ln(1 - \alpha_L) \qquad \forall t \tag{3.11}$$

Specialty Care: ($\ell = S$) In each period there are two kinds of patients who visit the infectious diseases specialty, patients under monitoring and patients under treatment, given bu, M_{r,t^i} and $D^i_{r,t}$. Patients of health state i who are under monitoring and treatment visit the healthcare system during a given period with frequency ϕ^i_M and ϕ^i_D respectively. Therefore, $\lambda_S = \sum_{i\in\mathcal{I}-\{13\},r\in\mathcal{R}}\left(M^i_{r,t}\phi^i_M + D^i_{r,t}\phi^i_D\right)$ and $\mu_S = \min_k \{n_{k,S}A_{k,S}/y_{k,S}\}$. Substituting these in inequality (3.9), we get the service level constraint at the infectious diseases specialty as,

$$\sum_{i\in\mathcal{I}-\{13\},r\in\mathcal{R}}\left(M^i_{r,t}\phi^i_M + D^i_{r,t}\phi^i_D\right) = \min_k \{n_{k,S}A_{k,S}/y_{k,S}\} + \frac{1}{\tau_S}\ln(1-\alpha_S) \qquad \forall t \tag{3.12}$$

3.3.3 Planning Problem

Using the above described objective function, system dynamics, and organizational constraints, the planning problem faced by the GLA station can be formulated as the following nonlinear mixed integer program, which we describe as the QALY Maximizing Planning Problem (QMPP).

(QMPP)

$$\sum_{i\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i\left(U^i_{r,t} + W^i_{r,t} + E^i_{r,t} + M^i_{r,t} + D^i_{r,t}\right)$$

subject to

$$(3.1) - (3.12)$$

$$0 \le S_{r,t} \le 1 \qquad \forall r,t \tag{3.13}$$

$$U^i_{r,t}, W^i_{r,t}, R^i_{r,t}, R^i_{r,t}, E^i_{r,t}, M^i_{r,t}, D^i_{r,t},$$

$$ID^i_{r,t}, I^i_{r,t}, IM^i_{r,t}, IM^i_{r,t} \in \mathbb{R}_+ \qquad \forall r,i,t \tag{3.14}$$

$$n_{k,l} \in \mathbb{N}_+ \qquad \forall k,\ell \tag{3.15}$$

Observe that the QMPP contains a knapsack problem defined by constraints (3.8). Thus, we need to solve instances of an NP-complete problem and it may not be always possible to solve real sized problems to optimality. We verified this in our computational experiments in §3.5. Consequently, to solve this problem, we elected to develop effective heuristics that are both computationally tractable and managerially intuitive. We also develop relaxations to the problem to obtain an upper bound on the objective function that is used to evaluate the performance of the heuristics. If we replace $\alpha\beta S_{r,t}U^i_{r,t}$ with $V^i_{r,t}$ in constraints (3.1), (3.2), (3.8), (3.10) of the QMPP and add the definitional constraint $V^i_{r,t} = \alpha\beta S^i_{r,t}U^i_{r,t}\ \forall r, i, t$, then the QMPP can be transformed into the following integer bilinear program QMPPB. This will be useful in developing a tight upper bound for the QMPP.

(QMPPB)

$$\sum_{i\in\mathcal{I}, r\in\mathcal{R}, t\in\mathcal{M}_\tau, \tau\in[T]} q^i \left(U^i_{r,t} + W^i_{r,t} + E^i_{r,t} + M^i_{r,t} + D^i_{r,t}\right)$$

subject to,

$$(3.3) - (3.7) \text{ and } (3.11) - (3.15) \text{ and}$$

$$U^i_{r,t+1} = \left(\sum_{j\in\mathcal{I}_w} \theta^{i,j}_{r,untreat}(1 - \alpha\beta V^i_{r,t}\right) + N^i_{r,t+1} + R^0_{r,t}\theta^{0,i}_{r,untreat} \qquad \forall r, i, t \tag{3.16}$$

$$\begin{aligned} W^i_{r,t+1} = {} & \sum_{j\in\mathcal{I}} W^j_{r,t}\theta^{j,i}_{r,untreat} + \sum_{j\in\mathcal{I}_w} \alpha\beta V^i_{r,t}\theta^{j,i}_{r,untreat} + \sum_{j\in\mathcal{I}_o} U^j_{r,t}\theta^{j,i}_{r,untreat} \\ & - \sum_{j\in\mathcal{I}} R^j_{r,t}\theta^{j,i}_{r,untreat} \qquad \forall r, i, t \end{aligned} \tag{3.17}$$

$$\begin{aligned} & \sum_{i\in\mathcal{I}_w, r\in\mathcal{R}, t\in\mathcal{M}_\tau} CS^i\alpha\beta V^i_{r,t} + \sum_{i\in\mathcal{I}_o, r\in\mathcal{R}, t\in\mathcal{M}_\tau} CS^i Ur,t^i \\ & \quad + \sum_{i\in\mathcal{I}_o, r\in\mathcal{R}, t\in\mathcal{M}_\tau, X\in\mathcal{X}} + \sum_{\ell\in\mathcal{L}, k\in\mathcal{W}, t\in\mathcal{M}\tau} n_{k,l}w_k \le B(\tau) \qquad \forall\tau \end{aligned} \tag{3.18}$$

$$\begin{aligned} & \sum_{i\in\mathcal{I}_w, r\in\mathcal{R}} \alpha\beta V^i_{r,t} + \sum_{i\in\mathcal{I}_o-\{13\}, r\in\mathcal{R}} U^i_{r,t} \le \min_k \left\{n_{k,P}A_{k,P}/y_{k,P}\right\} \\ & \qquad + \frac{1}{\tau_P}\ln\left(1 - \alpha_P\right) \qquad \forall t \end{aligned} \tag{3.19}$$

$$V^i_{r,t} = \alpha\beta S^i_{r,t} U^i_{r,t} \qquad \forall r, i, t \tag{3.20}$$

$$V^i_{r,t} \in \mathbb{R}_+ \qquad \forall r, i, t \tag{3.21}$$

Observe that in the integer bilinear program QMPPB, all the nonlinearity in the problem is now captured by bilinear constraints (3.20)

Proposition 6. *The objective function of the QMPPB can be written as*

$$K_0 + \sum_{i\in\mathcal{I}, r\in\mathcal{R}, t\in\mathcal{M}_\tau, \tau\in[T]} \pi^i_{r,t} D^i_{r,t} \tag{3.22}$$

Where, K_0 and $\pi^i_{r,t} = f\left(\theta^{j,i}_{r,treat}, \theta^{j,i}_{r,untreat}, q^i, t\right)$ are constants.

All proofs are provided in the Appendix. Proposition 6 implies that the QALYs in the system cannot be maximized by increasing the screening rate alone, as advocated by both the risk based and routine screening policies, unless that increase can be translated to patients treated. This is consistent with observations in population level studies (Long et al. 2010). However, the number of patients treated is often constrained by the budgetary and capacity constraints. Thus, the focus should be on determining how many patients can be optimally treated and this in turn should be used to determine the screening rates. This is accomplished by the QMPPB. Let $\underline{U}^i_{r,t}$ be a lower bound and $\bar{U}^i_{r,t}$ be an upper bound on $U^i_{r,t}$. The computations of these bounds are described in the appendix. The following proposition helps in reducing the complexity of the search space for heuristics to solve the QMPPB.

Proposition 7. *The screening rate is bounded by the following two inequalities:*

$$\sum_{r\in\mathcal{R}, t\in\mathcal{M}_\tau} \sigma_{r,t} S_{r,t} \leq B(\tau) - K_\tau - \sum_{i\in\mathcal{I}-\{13\}, r\in\mathcal{R}, t\in\mathcal{M}_\tau} \rho^i D^i_{r,t} \qquad \forall \tau$$

$$\sum_{i\in\mathcal{I}_w, r\in\mathcal{R}, t\in\mathcal{M}_\tau} \alpha\beta S_{r,t} \bar{U}^i_{r,t} \geq \sum_{i\in\mathcal{I}-\{13\}, r\in\mathcal{R}, t\in\mathcal{M}_\tau} D^i_{r,t} - \sum_{r\in\mathcal{R}, t\in\mathcal{M}_\tau} \bar{U}^i_{r,t} \tag{3.23}$$

Where, K_τ, ρ^i, $\sigma_{r,t}$ are given by,

$$\begin{aligned} K_\tau = \sum_{k \in \mathcal{W}, t \in \mathcal{M}_\tau} & \left\{ \sum_{i \in \mathcal{I}_o} \left(\frac{w_k y_k}{A_{k,P}} \right) \underline{U}^i_{r,t} - \left(\frac{w_k y_k}{A_{k,P} \tau_P} \right) \ln \left(1 - \alpha_P \right) \right. \\ & \left. - \left(\frac{w_k y_k}{A_{k,S} \tau_S} \right) \ln \left(1 - \alpha_S \right) \right\} + \sum_{i \in \mathcal{I}, r \in \mathcal{R}, t \in \mathcal{M}_\tau} C^i_U, \underline{U}^i_{r,t} + \sum_{i \in \mathcal{I}_o, r \in \mathcal{R}, t \in \mathcal{M}_\tau} CS^i \underline{U}^i_{r,t} \end{aligned} \tag{3.24}$$

$$\rho^i = C^i_D + \sum_{k \in \mathcal{W}} \left(\frac{w_k y_k}{A_{k,P} \tau_P} \right) \phi^i_D \qquad \forall i \tag{3.25}$$

$$\sigma_{r,t} = \sum_{i \in \mathcal{I}_w} \left\{ \sum_{k \in \mathcal{W}} \left(\frac{w_k y_k}{A_{k,P}} \right) \underline{U}^i_{r,t} + CS^i \underline{U}^i_{r,t} \right\} \qquad \forall r, t \tag{3.26}$$

Further, for a stationary screening policy for which $S_{r,t} = S_r$ $\forall t$, $S_r \leq \left(B(\tau) - K_\tau \right) / \sum_{t \in \mathcal{M}_\tau} \sigma_{r,t}$

Note from Proposition 7 that for a given screening rate, the total number of patients that can be treated is bounded by (1) the residual budget left over for treatment after the screening, staffing, and the patient state costs and (2) the number of screened asymptomatic patients who test positive and symptomatic patients being treated. Further, the total number of patients who actually are treated will be determined by whichever of these two conditions becomes tight. Given that typically budgets are scarce and there is a large population of patients, it is likely that the budget constraint would be tighter. This implies that while setting screening rates, one has to understand budgets and its implications on treatment. This is consistent with the public health literature (Martin et al., 2010).

3.3.4 Relaxations and Upper Bounds

To develop an upper bound on the QMPPB, we replace bilinear constraints (3.20) by convex over and under estimators of the bilinear terms using the approach proposed by McCormick (1976).

Let $\bar{U}^i_{r,t}$ and $\underline{U}^i_{r,t}$ represent the upper and lower bound on the variable $U^i_{r,t}$ respectively. Then it follows from (3.20) that,

$$V^i_{r,t} \geq \alpha\beta S^i_{r,t}\underline{U}^i_{r,t} \qquad \forall r, i, t \tag{3.27}$$

$$V^i_{r,t} \leq \alpha\beta S^i_{r,t}\bar{U}^i_{r,t} \qquad \forall r, i, t \tag{3.28}$$

Note that, $\alpha\beta S^i_{r,t}\underline{U}^i_{r,t} \leq \alpha\beta S^i_{r,t}U^i_{r,t} \leq \alpha\beta S^i_{r,t}\bar{U}^i_{r,t}$ and $0 \leq S_{r,t} \leq 1\ \forall r, i, t$. Then $\alpha\beta S_{r,t}\bar{U}^i_{r,t} + \alpha\beta U^i_{r,t} - \alpha\beta S^i_{r,t}\bar{U}^i_{r,t} = (S_{r,t} - 1)\alpha\beta\bar{U}^i_{r,t} + \alpha\beta U^i_{r,t} \leq (S_{r,t} - 1)\alpha\beta U^i_{r,t} + \alpha\beta S^i_{r,t} = S_{r,t}\alpha\beta S^i_{r,t} = V^i_{r,t}$. Thus,

$$V^i_{r,t} \geq \alpha\beta S_{r,t}\bar{U}^i_{r,t} + \alpha\beta U^i_{r,t} - \alpha\beta\bar{U}^i_{r,t} \qquad \forall r, i, t \tag{3.29}$$

Similarly, $\alpha\beta S_{r,t}\underline{U}^i_{r,t} + \alpha\beta U^i_{r,t} - \alpha\beta S^i_{r,t}\underline{U}^i_{r,t} = (S_{r,t} - 1)\alpha\beta\underline{U}^i_{r,t} + \alpha\beta U^i_{r,t} \leq (S_{r,t} - 1)\alpha\beta U^i_{r,t} + \alpha\beta S^i_{r,t} = S_{r,t}\alpha\beta S^i_{r,t} = V^i_{r,t}$. Thus,

$$V^i_{r,t} \leq \alpha\beta S^i_{r,t}\underline{U}^i_{r,t} + \alpha\beta\underline{U}^i_{r,t} - \alpha\beta S^i_{r,t}\underline{U}^i_{r,t} \qquad \forall r, i, t \tag{3.30}$$

Observe that constraints (3.27)-(3.30) provide a linear relaxation to bilinear constraints (3.20). This substitution reduces this problem to a linear mixed integer program that can now be solved to optimality using commercial solver such as the GUROBI solver (Gurobi Optimization, 2015). We call this formulation the RQMPPB and note that the optimal solution to the RQMPPB provides an upper bound to the QMPPB and consequently the QMPP.

The quality of this upper bound strongly depends on the bounds of $U^i_{r,t}$. A recent improvement to the McCormick relaxation is introduced by Wicaksono and Karimi (2008). We adapt this technique to do an *ab initio* partitioning on $U^i_{r,t}$ apply a set of under and over estimators to each partition, and introduce a logical constraint to limit the partitioned variable to one active partition. To achieve this, let $U^i_{r,t}$ be separated into m equally spaced partitions as $\underline{U}^i_{r,t} = a^i_{r,t}(1) < \cdots < a^i_{r,t}(m) < a^i_{r,t}(m+1) = \bar{U}^i_{r,t}$. The choice of parameter m is based on comparing the reduction in the value of the bound with the increased time it takes to compute the bound when m is incremented by one starting with $m = 1$ and is described in the appendix.

Define binary variable $\xi^i_{r,t}(m)$ so that $\xi^i_{r,t}(m) = 1$ if $U^i_{r,t} \in \left[a_{r,t}(m)^i, a^i_{r,t}(m+1)\right]$ and $\xi^i_{r,t}(m) = 0$ otherwise. This leads to the following constraints:

$$U^i_{r,t} \geq a^i_{r,t}(m)\xi^i_{r,t}(m) + \underline{U}^i_{r,t}[1 - \xi^i_{r,t}(m)] \qquad \forall r, i, t, m \tag{3.31}$$

$$U^i_{r,t} \leq a^i_{r,t}(m+1)\xi^i_{r,t}(m) + \bar{U}^i_{r,t}[1 - \xi^i_{r,t}(m)] \qquad \forall r, i, t, m \tag{3.32}$$

$$\sum_{m=1}^{M} \xi^i_{r,t}(m) = 1 \qquad \forall r, i, t \tag{3.33}$$

$$\xi^i_{r,t}(m) \in \{0, 1\} \tag{3.34}$$

Next, we introduce constraints of the type (3.27)-(3.30) for each partition by replacing $\bar{U}^i_{r,t}$ with $a^i_{r,t}(m+1)$ and $\underline{U}^i_{r,t}$ with $a^i_{r,t}(m)$. Depending on $\xi^i_{r,t}(m)$, the appropriate set of constraints wold be activated, thus providing tight relaxation to the bilinear terms. This leads to the following constraints,

$$V^i_{r,t} \geq \alpha\beta S_{r,t} a^i_{r,t}(m) - K[1 - \xi^i_{r,t}(m)] \qquad \forall r, i, t, m \tag{3.35}$$

$$V^i_{r,t} \leq \alpha\beta S_{r,t} a^i_{r,t}(m+1) + K[1 - \xi^i_{r,t}(m)] \qquad \forall r, i, t, m \tag{3.36}$$

$$V^i_{r,t} \geq \alpha\beta S_{r,t} a^i_{r,t}(m+1) + \alpha\beta U^i_{r,t} - \alpha\beta a^i_{r,t}(m+1) - K[1 - \xi^i_{r,t}(m)] \qquad \forall r, i, t, m \tag{3.37}$$

$$V^i_{r,t} \leq \alpha\beta S_{r,t} a^i_{r,t}(m) + \alpha\beta U^i_{r,t} - \alpha\beta a^i_{r,t}(m) + K[1 - \xi^i_{r,t}(m)] \qquad \forall r, i, t, m \tag{3.38}$$

The value of parameter K is set sufficiently large to deactivate these constraints if $U^i_{r,t}$ does not belong to that particular partition. To provide a tighter upper bound on the QMPPB, we solve the RQMPPB by replacing (3.27)-(3.30) with (3.35)-(3.38). The performance of this bound is evaluated in §3.5

3.3.5 Heuristics and Lower Bounds

In this section, we discuss several possible heuristic solution methods to the QMPPB that correspond to potential implementation policies at the GLA station. They can broadly be classified as fixed staffing heuristics and variable staffing heuristics.

Fixed Staffing Heuristics Here, we do not optimize over the staffing variables $n_{k,\ell}$ $\forall k, \ell$ and these are set to existing levels corresponding to the risk based screening policy. In this case, QMPPB reduces to a continuous bilinear program. We then develop two heuristics depending on how the screening rate varies over time. In the first heuristic, we add constraint $S_{r,t} = S_r$ $\forall r, t$ to ensure that the recommended screening policy is stationary. Although apparently restrictive, it is easy to implement and thus was appealing. To solve the resulting problem we iteratively narrow down on the optimal stationary fixed screening using the search algorithm described in the Appendix. Note that this algorithm is quite simple to implement because evaluation of the QMPPB given the screening rates is now a linear program and can be solved very effectively using several commercially available solvers such as the GUROBI solver. Further, Proposition 7 enables us to reduce the solution space of this algorithm. We refer to this heuristic as the Fixed Staffing Stationary Screening (FSSS) heuristic.

In the second heuristic, we allow the screening rate to vary over time so that the resulting screening policy is non-stationary. The resulting problem reduces to a continuous bilinear program which is solved by using the generalized reduced gradient algorithm (Abadie and Carpentier, 1969). This algorithm has been shown to be very effective for large sparse dynamic nonlinear optimization problems (Drud, 1985). We refer to this heuristic as the Fixed Staffing Non-stationary Screening (FSNS) heuristic. Clearly this heuristic is less restrictive than the FSSS is and hence can be expected to perform better.

Variable Staffing Heuristics. Next, we describe two heuristics, where we allow the staffing levels to change and again consider either stationary or non-stationary screening rates. We refer to these as the Variable Staffing Stationary Screening (VSSS) and the Variable Staffing Non-stationary Screening (VSNS) heuristic, respectively. The solution procedure for the VSSS heuristic is very similar to that of the FSSS heuristic, with the key difference being that the evaluation of the QMPPB for a given screening rate in the search algorithm would now require solving a mixed integer program. Although this potentially can be more complicated, we found that the GUROBI solved this problem very effectively. The solution to the VSNS heuristic is complicated as it involves solving a nonlinear mixed integer program. We employ the combined penalty and outer approximation method (Vishwanathan and Grossman 1990) to solve this problem. Given that we can optimize both staffing levels and the screening rates in the variable staffing heuristics, we expect both of them to outperform the corresponding fixed staffing heuristics. However, the magnitude of the gap between these heuristics is not apparent. Similarly, whether the VSSS outperforms the FSNS or vice versa is not obvious a priori. We investigate these issues in the computational experiments in §3.5

Finally, observe that the QMPPB is not jointly convex in the decision variables. Thus, this sequential approach in the FSSS and the VSSS provides a feasible but not necessarily an optimal solution. Similarly, given the complexity of the QMPPB, the algorithms used to execute the FSNS and VSNS provide feasible but not optimal solutions.

3.4 Data Collection and Model Validation

The data required for our model can be divided into two broad categories. The first category includes operational data concerning costs, budgets, incoming patient characteristics, time required for various activities, time available, and service level parameters. These data are specific to the GLA station and were collected from a variety of sources including direct observation, administrative databases, and clinical studies. The second category includes clinical data on visit frequency under HIV care, the quality of life estimates for HIV patients in different health states, and treatment decisions. We use published estimates for these parameters from the existing clinical literature that are more broadly applicable. Below we describe each of these categories in greater detail. We then use the data to validate our model both in the context of the literature and the GLA station.

3.4.1 Operational Data

Costs Primary drivers for variable cost in our model are cost of HIV screening cost (CS^i) , system state cost (C^i_X) i per patient, and wages (w_k). The screening cost CS^i consists of the material cost of screening. The screening cost per patient was estimated to be \$80. The system state cost per patient C^i_X is composed of several components. Therefore, its estimation is more involved and discussed in the Appendix. Because the staffing levels are endogenous to the model, the other relevant cost component are the wages paid to the healthcare workers of different types (w_k). At the GLA station, these costs are fixed and do not vary based on the patient load. These are shown in the appendix.

Budget The VHA allocates the budget to the GLA station annually, and this budget does not carry over to the next year. To provide a more stable and a long range plan, we conduct our analysis for a period of two years, where the budget for year is given by $B(\tau) \in \{1, 2\}$ Note that our model can be easily extended for $\tau > 2$ without any changes to the methodology by the appropriate choice of T , where $\tau \in [T] = \{1, 2, \ldots, T\}$ This is described in the appendix. However, extending the model beyond two years was not realistic in our application context because there was significant uncertainty in the costs of screening and treatment, the population of veterans that would be served at this station, and the incidence and prevalence rates. To incorporate the uncertainty in these parameters, the model can be solved every year with a two-year horizon using updated parameters.

Because of various complexities in estimation, the annual GLA station budget was not broken down to the level for HIV related activities, which is the focus of our analysis. Therefore, we imputed a budgetary range $[\underline{B}(\tau), \bar{B}_\tau]$ using the risk based screening policy currently followed at VHA (i.e., $S_{1,t} = 1$ and $S_{2_t} = 0 \forall t$. The lower bound of this range corresponds to the smallest annual budget at which the risk based screening policy is feasible. The upper bound corresponds to the smallest value of the annual budget at which no further gains in QALYs can be accrued from the risk based screening policy. This approach is formalized in the budget imputation algorithm provided in the Appendix. We conduct our analysis on all the proposed policies within this budgetary range.

Incoming Patient Characteristics Let N_t denote the number of new patients entering the station in time period t and $\hat{p}_r^i$ be the fraction of these patients in risk category r and health state i. The number of new patients in each risk category and health state in each period who enter the station is thus given by $N_{r,t}^t = N_t \hat{p}_r^i$. To estimate N_t we calculated the mean of historical data of total incoming patients over the past 12 months. The variation around the mean was negligible and we did not detect any temporal trends (such as increasing or decreasing over time) for the number of new patients. The parameter $\hat{p}_r^i$ is the proportion of patients in each risk and CD4 category. We calculate $\hat{p}_r^0 = (1 - prev_r)$, where prevalence rate $(prev_r)$ is estimated by Paltiel et al. (2005) and shown in the appendix. The proportion of patients who are infected $(prev_r)$ is further divided into different CD4 counts in a fraction estimated for the VHA by Gandhi et al. (2007), thus determining $\hat{p}_r^i$, $\forall i \neq 0$. We report this in the appendix. We were provided with

U_1 the total number of patients currently enrolled at the GLA station. Thus, the number of unscreened patients in each risk category and each health state would be given by $U^i_{r,1} = \hat{p}^i_r U_1$. The fraction of patients who visit a healthcare facility for non-HIV related reasons was estimated by dividing the total number of unique patients who visited the inpatient or the outpatient facilities for non-HIV related reasons by the total number of patients registered in the station. Using this approach, we estimated $\alpha = 0.5$. The proportion of patients who accept screening β was assumed to be 50% based on prior studies (Goetz et al., 2008b).

Time Required, Time Available, and Service Level Parameters. To estimate $y_{k,l}$, the time required per patient of healthcare worker of type k at location l, we used an observational time and motion study conducted in the emergency department in the VA West Los Angeles Medical Center within the GLA station (Gidwani et al., 2012). These, data shown in appendix, were validated against other published estimates (Silva et al., 2007). We note that these times would be very similar for other care settings in the station such as the primary care clinics, inpatient department, and outpatient department.

The total time available at each resource at each location per month, $A_{k,\ell}$ for activities associated with the routine HIV screening program was based on estimates from the GLA station. It took into account that healthcare workers need to devote time to other clinical and administrative activities as well. These estimates are shown in the appendix.

Lastly, it was expected that at least 95% of all patients should be processed at each location within a period of one month. Thus, $\tau_\ell = 1, \alpha_\ell = 0.95$.

3.4.2 Clinical Data

Visit Frequency Under HIV Care The outpatient visit frequency for VHA was not directly available. We used published estimates by Schackman et al. (2006) for the frequency of outpatient visit under monitoring (ϕ_M^i) under treatment (ϕ_D^i). This is reported in the appendix.

Quality of Life (QOL) Utilities. The QOL utilities were drawn from Freedberg et al. (1998) and Mauskopf et al. (2005). These are summarized in Table 3.3 and more details are provided in the appendix. Here, it was assumed that the health related quality of life utilities (q^i) are directly associated with the underlying health state represented by the CD4 count category and OI infection status rather than on the treatment status *per se*. This is reasonable because the effect of treatment is eventually reflected in patients being in better health states and hence enjoying a higher QOL utility.

Table 3.3: QOL Weights

Health state index (i)	QOL weight q^i	Health state index (i)	QOL weight q^i
0	1	5	0.81
1	0.94	6	0.79
2	0.94	7-12	0.60
3	0.94	13	0
4	0.87		

Treatment Decision The treatment policy at the GLA station was to initiate patients having CD4 cell count below 350 cells/mm^3 and patients with opportunistic infection irrespective of their CD4 count on treatment and retain the rest on monitoring. From Table 3.1, this implies that $z^i = 0$ for $i = \{0, 1, 2\}$ and $z^i = 1$ otherwise.

3.4.3 Model Validation

In this section, we conduct analyses to validate the model in the context of the literature and the GLA station. To ensure an unbiased comparison with the literature (Paltiel et al., 2005), we removed all the organizational constraints in the model so that it reduces to a pure disease progression and treatment model as considered by these papers. Bishai et al. (2007) calculate total QALYs gained from treatment over no treatment for HIV positive patients. We used their treatment regimen in our model and found that the total QALYs gained was comparable to their work. Paltiel et al. (2005) calculates the amount spent per QALY gained from going from no treatment to treatment under various screening policies and found that this varied between \$63,000 and \$113,000 spent per QALY gained. We also used our model to calculate the amount spent per QALY gained for the different policies in Paltiel et al. (2005) and found it to be similar, ranging from \$61,000 to \$111,000 spent per QALY gained. This validates that our disease progression and treatment model is consistent with the literature.

In the context of the GLA station, we considered the entire model and the current risk based screening policy. We found that the model estimates on the number of people at each disease state, location, and time period were within 2% of the actual numbers at the GLA station. We also used the resulting arrival rate λ_ℓ and service rate μ_ℓ at location $\ell \in \{P, L, S\}$ to estimate $\bar{W}_\ell = 1/(\mu_\ell - \lambda_\ell)$, the average wait times at each location for a given time period under the $M/M/1$

queuing model assumption used in deriving the service level constraints (Kleinrock 1975). We found these estimates were within 5% of the actual average wait times for the corresponding locations and time period at the GLA station. This supported the rationale for using the $M/M/1$ queuing model in developing the service level constraints. These analyses also validate that our model effectively captures the operating environment at the GLA station and is a necessary step to provide confidence in the policy analysis described next.

3.5 Policy Analysis

In this section, we evaluate several policies for screening, testing, and care within the framework of our model. We start with analyzing the risk based screening policy that had been the standard of care at the VHA when we started this work. We then evaluate the impact of the routine screening policy under consideration and also assess the performance of the heuristics described §3.3.5.

Recollect from §3.4.1 that the annual budget expenditure required for HIV screening, treatment, and monitoring was not directly available. Therefore, we used the budget imputation algorithm provided in the appendix to first to impute the budget range $[\underline{B}(\tau), \bar{B}(\tau)]$ for the risk based screening policy in which $S_{1,t} = 1$ and $S_{2,t} = 0 \; \forall t$. Here, we found that $\underline{B}(\tau) = \$10$,million and $\bar{B}(\tau) = \$20$ million for $\tau = 1$ and 2. This implies that at least \$10 million is needed annually to implement the risk based screening program and any budget allocation over \$20 million will not improve the efficacy of this program further. We also used this algorithm to find that an annual budget of \$35 million was required to implement the routine screening policy in which $S_{r,t} = 1 \forall r, t$ Although this estimate was instructive, this level of funding may not be available in the foreseeable future. Therefore, the emphasis was in improving upon the risk based policy but within the current budgetary range of \$10 to \$20 million. To perform this analysis and simplify the exposition, we conducted all our subsequent analysis at three budget levels, low, medium, and high corresponding to \$14, \$16, and \$19 million, respectively. We tried to solve the QMPP for these budget values using leading commercial solvers for nonlinear mixed integer programs such as BARON and DICOPT using the NEOS server (Dolan et al., 2002). However, in all cases, these solvers could not even generate feasible solutions after more than 40 hours of computation, and the runs were aborted. This provides validation for developing bounds and heuristics to address this problem.

3.5.1 Performance of Heuristic Policies

We solved the FSSS, FSNS, VSSS, and VSNS using the approaches described in §3.3.5 and then calculated the QALYs gained from these four heuristic policies. We used the technique described in §3.3.4 to compute the upper bounds for each of these budgetary levels. The computations for the risk based screening policy, the routine screening policy, FSSS, VSSS, and upper bounds were executed with GUROBI, a general purpose LP/MIP solver using the NEOS server. The computations for the FSNS and the VSNS were implemented with DICOPT using the NEOS server. All heuristics were solved in a few seconds, whereas each computation of the upper bound took at most three hours. Note that in computing the upper bounds for the fixed staffing heuristics FSSS and FSNS, we fixed the staffing levels at the current levels at the GLA station. This ensured that these heuristics were being fairly compared to an upper bound to the fixed staffing problem. We measured the performance of the heuristics using a percentage gap defined as the difference between QALYs gained from the upper bound and those gained from the heuristic policy expressed as a percentage of the QALYS gained from the upper bound. In all cases, QALYs gained were calculated with the base case of no screening. Table 3.4 summarizes the gaps for the four heuristics across the three budgetary levels.

The percentage gaps described in Table 3.4 indicate that all the heuristics perform very well. In particular, the average gap across these heuristics is 1.95% and ranges from 0.08% to 5.15%. In general, for the fixed staffing heuristics, the gaps increase as the budget level increases. This is because the upper bounds increase at a greater rate than the heuristic solution does. The rate of growth of the heuristic solution is limited as the benefits from choosing the optimal screening rates at higher budget levels saturate because of fixed staffing in which more patients cannot be treated because of capacity and service level constraints. Conversely, for the variable staffing heuristics, the gaps decline as the heuristic solution increases at a greater rate than the upper bound. This is because variable staffing allows more effective allocation of staff at the higher budget levels to treatment and allows more screened patients who are diagnosed with HIV to be treated optimally and this improves the overall performance of the heuristics.

We also conducted sensitivity analysis to understand how parameters such as time available for HIV screening programs; service level parameters and the costs of wages, screening, and treatment affect these gaps for the heuristics. To perform this analysis, we first set the budget level to $16 million and changed each of these parameters one at a time from their base level by 30% to 30% in increments of 10%. We then calculated the gap for each heuristic and the appropriate change in the gap from the baseline reported in Table 3.4. Across all heuristics and range of values of these parameters, we found the average change in gaps was 3.3%, and this varied from 0.8% to 7.2%. This shows that these heuristics and the upper bounds are robust across a wide range of parameter values.

Table 3.4: Percentage gap of heuristics and percentage improvement from current practice.

Heuristic	% gap	% improvement
	Budget level: Low	
FSSS	0	20.18
FSNS	0	20.21
VSSS	4.32	283.90
VSNS	7.05	305.30
	Budget level: Medium	
FSSS	0.08	23.39
FSNS	0.2	24.13
VSSS	3.25	66.47
VSNS	5.15	69.69
	Budget level: High	
FSSS	1.27	38.80
FSNS	1.33	40.15
VSSS	0.48	41.53
VSNS	3.9	42.94

3.6 Improvements from Risk Based Screening

We computed the QALYs accrued at these budget levels for the current risk based screening policy. We used this to calculate the percentage improvement of the heuristics from the risk based screening policy expressed as a percent of the risk based screening policy solution. The results, summarized in Table 3.4, lead to the following observations. First, irrespective of the budget level, improvements from risk based screening increased as we go from the FSSS to the FSNS to the VSSS and finally to the VSNS heuristic. In particular, the most improvement is obtained from the VSNS because this policy synchronizes the screening decision with the staffing decision. This is important since it is ineffective to screen as many patients as possible and not have sufficient funding to treat them as necessary. Rather, it is critical to screen as many patients that can be optimally treated because the benefits arise only from treatment and not screening. This was shown in Proposition 1. This implies that one should first calculate how many people can be optimally treated and then use this to appropriately calculate the optimal screening rates. This approach is executed by the solution method of the VSNS. Second, note that the FSNS improves upon the FSSS by at most 3.47% and this is only 0.14% in the most realistic low budget scenario. This suggests that if staffing cannot be changed because of organizational reasons, then it is better to keep a stationary screening policy in the short term since it is easier to implement. However, if the long term goal is to accrue maximum benefit using the VSNS, the FSNS would be a good approach to allow the staff to get acclimatized to using non stationary screening rates prior to implementing the more radical changes associated with variable staffing. Third, the gains from varying staffing are more significant than those obtained by varying screening across any budget level. To see this, observe from Table 3.4 that the gains from going from fixed to variable staffing (i.e., FSSS to the VSSS or FSNS to the VSNS) are larger than the gains from stationary to non-stationary screening (i.e., FSSS to the FSNS or VSSS to the VSNS). Fourth, the benefit from variable screening is greater if staffing is allowed to change (i.e., the gains from VSNS-VSSS ¿ FSNS-FSSS). Finally, the greatest improvements from current practice occur in low budgets or resource constrained environments. This is because the optimization executing these policies ensures that screening and staffing rates are chosen in such a way that these scarce resources are used in the best possible manner.

Finally, we again conducted sensitivity analysis to study how the percentage improvement of the heuristics from the risk based screening policy change with model parameters such as time available for HIV screening programs, service level parameters and the costs of wages, screening, and treatment. To do so, we first set the budget level to \$16 million and changed each of these parameters one at a time from their base level by 30% to 30% in increments of 10%. In practice, such changes may be needed because of organizational requirements. As expected, the QALYs gains from all the heuristics declined as available time for HIV programs ($A_{k,\ell}$) and the service level parameter related wait time at location ℓ, τ_ℓ decreased. Similarly, the QALYs gained from the heuristics declined as the service level parameter related to the probability of meeting a wait time at location ℓ, α_ℓ , cost of wages, screening, and treatment increased. However, in all these cases, the relative gain from the benchmark risk based screening policy is increasing as the optimization inherent in the heuristics allowed them to better cope with diminished resources, higher service level requirements, or increased costs. In addition, the previously described order of improvement from FSSS to FSNS to VSSS to VSNS was still preserved. This shows that the comparative performance of the heuristics across a wide range of parameters is quite consistent, and they are better in coping with changes in these parameters values than the risk based screening policy.

3.6.1 Screening Rates and Staffing Allocation

We studied how the screening rates and staffing allocation vary for each of these policies at different budget levels. We start by discussing the screening rates across the policies. Here, we found at low budget levels, the screening rates of the variable staffing heuristics were higher than those of the fixed staffing heuristics. This is because fixing the staffing levels to those of the risk based screening policy resulted in a large portion of the budget being committed, thereby leaving little flexibility to increase screening rates. On the other hand, at higher levels of budget, the screening rates of the fixed staffing heuristics are now higher than the variable staffing heuristics. This is because once the staffing levels are fixed, the only way to utilize the additional budget and improve the solution is to increase screening rates. In contrast, the variable staffing heuristics balances the screening rates and staffing levels with the available budget in both these budget scenarios and thus yields a better solution. We also analyzed how screening rates vary over time in the non-stationary screening rate policies (i.e., FSNS and VSNS). Observe from Figure 3.2 that in both the FSNS and VSNS policies, screening rates ramp up, saturate at a stable level, and ramp down across a budget horizon. The ramp up occurs because there is a large pool of unscreened patients at the start of the horizon. Screening these patients at high rates would require numerous staff at screening and thus less staff would be available at treatment. This would lead to an undesirable outcome of screening patients without treating them. To prevent this from happening, both these policies ramp up screening rates to spread the workload over time with fewer staff at screening so that the remaining staff can be effectively utilized in treatment. This ramp up continues until the system reaches the desired balance between screening and staffing; at this point, the screening rate stabilizes. This screening rate is maintained until the time horizon for the current budget cycle draws to a close. At this point, the screening rates ramp down and more resources are focused on the treatment of screened patients to make sure that screened patients not treated in this horizon do not congest treatment in the next horizon. This is important because residual budgets from the current cycle do not carry over to the next cycle.

Next, consider the staffing allocation between primary care (i.e., where screening is conducted) and specialty care (i.e., where treatment is conducted) across policies. This is summarized in Figure 3.3. From this figure it can be seen that more staff was allocated to primary care compared to specialty care in the fixed staffing heuristics, whose staffing levels are set to the current risk based screening policy. This follows as in the risk based screening policy, all high risk patients are screened without explicitly determining the staffing requirement for treatment. This leads to lower QALYs in the system because many people are screened but may not be effectively treated. Conversely, the variable staffing policies allocated more staff to specialty care than primary care. This ensured that the number of patients treated and the resulting system-wide QALYs are maximized since, as shown in Proposition 6, these are accrued from treatment and not from screening. Finally, we observed that the staffing level in variable staffing heuristics was actually lower than those in the fixed staffing heuristics. This was a direct consequence of optimizing the allocation between primary and specialty care in the variable staffing heuristics based on the number of patients that can be treated. This, in turn, reduced the staffing level needed at screening to a greater extent than the increase in staff needed at treatment.

To summarize, the policy analysis conducted in this section has led to many organizational implications at the GLA station. These are discussed next.

Figure 3.2: Screening rates over time for the non-stationary screening rate policies.

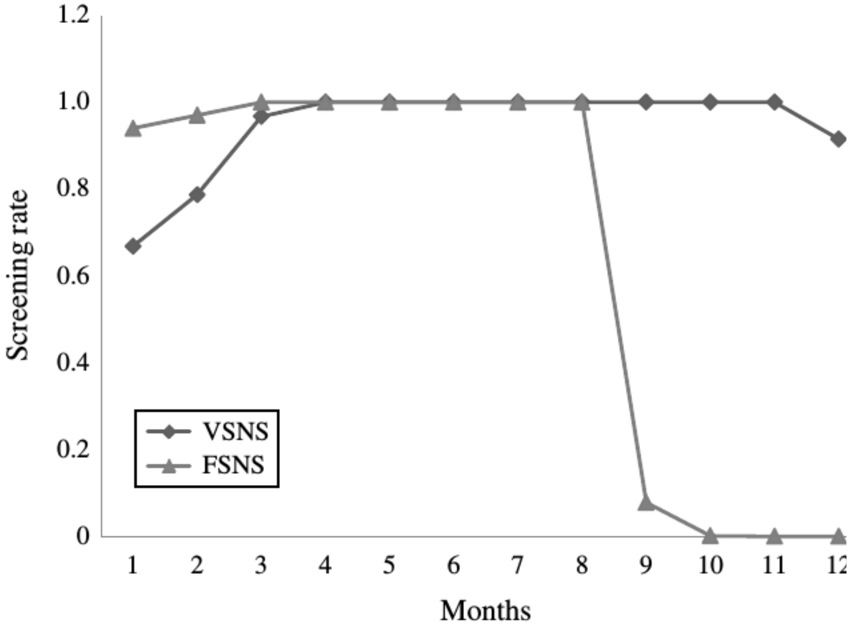
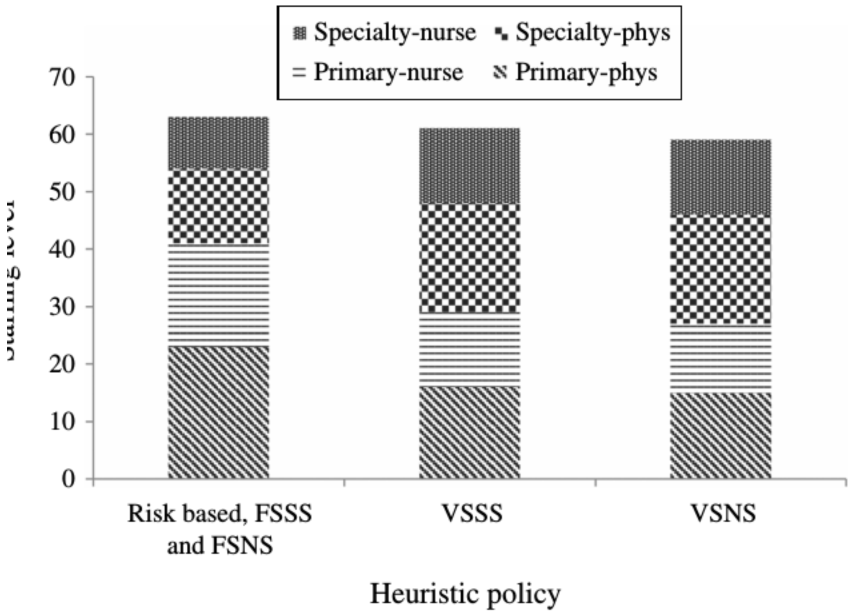


Figure 3.3: Staffing allocation across policies.



3.7 Application and Discussion

Several ideas developed in this paper have influenced decision making at the GLA station. A simplified version of the FSSS and the FSNS has been used to compute screening rates (Anaya et al., 2012). The rates ranged from 15% to 30% for the risk categories. These rates were considered to be reasonable and achievable. Further, they are consistent with research on HIV screening rates in other healthcare settings (Martin et al., 2010). The rates from the FSSS and the FSNS can be used to compute how many patients could be estimated to be present at the primary care, laboratory, and the infectious disease specialty over time. This information can then be used in constraint (3.8) to estimate the appropriate costs at different parts of the GLA station. This could provide valuable input for planning in future budgetary cycles. In addition, our methods show how these costs changed from the risk based screening policy to the FSSS and the FSNS. This provides an important justification in gaining the necessary funding in these budget cycles to implement these policies.

The implementation of expanded testing programs such as the FSSS and FSNS has led to early detection and early transfer to care for an increased number of patients. This in turn has resulted in better patient outcomes because they are identified at a stage of disease where the more serious manifestations of the illness are less common and when the response to therapy is better (Goetz and Rimland, 2011). The challenges in implementing these policies include educating the patients about the procedure and benefits of early testing, overcoming the reluctance of the providers to screen and prescribe these tests to patients they considered low risk or older and in stable monogamous relationships, training the staff at primary care to execute screening correctly, ensuring tests are conducted and information passed to care in a timely manner, and ensuring that patients are connected to care in an effective manner. Once patients are connected to care, it is important that there are sufficient updates of their health state information to ensure effective planning of staff for incoming patients in future periods. To ameliorate the impact of these challenges, the GLA station started implementation at its largest facility and used this learning to roll out to the whole station and other stations at the VHA (Goetz et al., 2011).

In addition, this work has had several managerial implications. It has shown that even though a policy such as routine based screening may be cost effective from a societal point of view, its implementation may not be feasible in an organization because of budgetary constraints. In particular, we show that at least a $15 million or 75% increase of annual budgetary outlays would be required to implement this policy from the risk based screening policy. This may not be possible at the GLA station because of the existing budgetary environment. Therefore, this provides the motivation to improve upon the risk based screening policy and we propose the FSSS, FSNS, VSSS, and the VSNS policies. Our analysis of these policies (summarized in Table 4) showed that optimizing the screening rate with existing staffing levels could increase the QALYs gained from risk based screening by 20% to 40%, or to 295 and 1,094 QALYs gained at the low and high budget levels, respectively. Further, in the low budget scenario, optimization of screening and staffing levels could increase QALYs gained from 245 for risk based screening5 to 995 or by over 300%. The approach we propose improves on risk based screening as it focuses on treatment, determines how many patients can be treated effectively, and then decides the appropriate screening rate. This is crucial because treatment determines the QALYs accrued in the system, in contrast to risk based screening where all high risk patients are screened without consideration of the staffing implications for treatment. In particular, the staffing implications of our variable staffing policies at the GLA station are more staff should be allocated to specialty care, lesser to the primary care, and this allocation in fact lowered total staff requirements. Although such staffing policies are harder to implement from an organizational perspective, we show this could result in significantly more gains, providing the management with the justification to consider these policies. Furthermore, we find that greatest benefit under variable staffing can be got by non-stationary screening. Here, it is beneficial to initially ramp up the screening rate to even the workload over time at treatment, allow this rate to stabilize, and finally ramp down toward the end of the budget cycle so that the remaining budget can be effectively used for treatment of patients. Finally, it is encouraging to note that the greatest gains can be achieved by these policies from risk based screening at the most realistic low budget scenario. In addition, the gains are increasing in order of FSSS to FSNS to VSSS to VSNS and this is independent of any budget scenario. Therefore our analysis provides direct justification for the GLA station to next consider the variable staffing policies (i.e., the VSSS and the VSNS) as the logical extension of the FSSS and the FSNS. Further, our method provides close to optimal staffing allocation and screening rates to successfully execute such variable staffing policies.

This work has the following limitations. First, our model does not account for the societal benefits of early screening by reducing transmission and ultimately prevalence rates. However, it is not possible to analytically estimate this reduction because it depends on individual behavior (i.e., whether one would take adequate precautions after being diagnosed) and if the people affected by this individual are a part of the VHA system. Therefore, we systematically reduced prevalence rates to calculate the impact on budgets and QALYs gained. The results summarized in the appendix show that even small reduction in prevalence rates could significantly lower budget requirements or increase QALYs gained. Second, we have assumed only two risk categories in determining screening rates and do not further stratify based on race and ethnicity because there are no clinical studies that can then be used to estimate transition rates between several health states. However, such divisions may increase the efficacy of our methods by early identification and treatment of certain patient groups. Third, several model parameters such as visit frequency, QOL utilities, incidence, and prevalence rates were estimated using clinical literature based on the general HIV population because they were not available specifically to the GLA station. To improve the performance of our methods, these parameters need to be updated as results from more current clinical studies become available or studies specific to the GLA station are conducted. Finally, our analysis is conducted at the station level for budgetary and staff allocation reasons. To keep this aggregate analysis tractable, we assumed a compartmental model with deterministic transitions between health states. However, this approach leads to a loss of granularity in terms of patient flows. Specifically, we do not consider the differences in cost and treatment effectiveness of individual patients in a particular health state. Further, we do not incorporate prioritization decisions that may be made within a health state due to presence of other health conditions of the patients such as heart disease, diabetes, or cancer. To consider these aspects in a shorter time horizon, one needs to consider a more detailed scheduling model with stochastic transition between disease states, and this is beyond the scope of our study.

In conclusion, we developed a model to address the screening and staffing decisions for HIV screening, testing, and care at the GLA station of the VHA. We applied this model to evaluate the risk based screening policy that was being used and also showed that the cost effective routine screening policy recommended by the CDC may not be feasible in this organizational context because of budgetary constraints. Therefore, we developed alternative fixed staffing policies within the framework of our model that are feasible and determined the relative improvement from using these policies from the risk based screening policy. We also developed managerial insights to better understand these policies and provided justification to the station administration to further extend and enhance their use by considering the variable staffing policies. This paper opens up several opportunities for future work. First, further work could improve the heuristic policies and the upper bound to reduce the sub-optimality gap. Second, this framework can also be used to evaluate HIV screening, testing, and care in other healthcare systems that have periodic patient follow up and in which residual budgets do not carry over to future periods (Petersen et al., 2007). In these settings, our existing modeling framework may have to be changed to include alternative objective functions, system dynamics, and organizational constraints. This could require development of different solution methods and bounds. Finally, a similar modeling framework can be used to assess the feasibility of other cost effective interventions (such as in tuberculosis and cardiac care) and if needed, develop alternative policies that improve current practice and are feasible from an organizational perspective.

3.8 Appendix

Estimation of Bounds on $U^i_{r,t}$

We describe the calculation of the lower bound $\underline{U}^i_{r,t}$ and the upper bound $\bar{U}^i_{r,t}$ on $U^i_{r,t}$. These parameters are used in Proposition 7 to reduce the search space of the search algorithms and are also important parameters in the method described in §3.3.4 used to develop upper bounds on the QMPPB. From Equation (3.1), we get

$$\begin{aligned}
U^i_{r,t+1} &= \left(\sum_{j \in \mathcal{I}_w} \theta^{j,i}_r (1 - \alpha\beta S_{r,t+1}) U^j_{r,t} \right) + N^i_{r,t} + R^0_{r,t}\theta^{0,i}_{r,untreat} \\
U^i_{r,t+1} &\geq \left(\sum_{j \in \mathcal{I}_w} \theta^{j,i}_r (1 - \alpha\beta S_{r,t+1}) U^j_{r,t} \right) + N^i_{r,t} \\
U^i_{r,t+1} &\geq \left(\sum_{j \in \mathcal{I}_w} \theta^{j,i}_r (1 - \alpha\beta) U^j_{r,t+1} \right) + N^i_{r,t} + R^0_{r,t}\theta^{0,i}_{r,untreat} \qquad \text{since, } S_{r,t} \leq 1
\end{aligned}$$

If we can find $\underline{U}^i_{r,t} \leq U^i_{r,t}$, then,

$$U^i_{r,t+1} \geq \left(\sum_{j \in \mathcal{I}_w} \theta^{j,i}_r (1 - \alpha\beta) \underline{U}^j_{r,t} \right) + N^i_{r,t}$$

Therefore, we get the recursive formula,

$$\underline{U}^j_{r,t+1} = \left(\sum_{j \in \mathcal{I}_w} \theta^{j,i}_r (1 - \alpha\beta) \underline{U}^j_{r,t} \right) + N^i_{r,t}$$

Also, $U^i_{r,1} = U_1 \hat{p}^i_{r,t}$ (both known numbers). Then $\underline{\mathrm{U}}^i_{r,1} = U^i_{r,1}$ and we recursively build the lower bounds.

Proof of Proposition 6

We first use induction on t to show the following equation holds. Let $K^i_{r,t}$ and $v^i_{r,t}$ be constants, then,

$$\begin{aligned}
U^i_{r,t} + W^i_{r,t} + E^i_{r,t} + M^i_{r,t} + D^i_{r,t} = K^i_{r,t} \\
+ \sum_{i \in \mathcal{I}, s \in \{1,2,\ldots,t\}} v^i_{r,s} D^i_{r,s}
\end{aligned} \qquad (3.39)$$

Observe that for $t = 1$ (3.39) is trivially true, since $U^i_{r,1}$ is a constant and $W^i_{r,1}$, $E^i_{r,1}$, $M^i_{r,1}$, $D^i_{r,1}$ are all zero. Next assume that, (3.39) is true for some t. We show that it holds for $t+1$. From the system dynamics equations (3.1)-(3.7), we get,

$$\begin{aligned} & U^i_{r,t+1} + W^i_{r,t+1} \\ & + E^i_{r,t+1} + M^i_{r,t+1} + D^i_{r,t+1} = \sum_{j\in\mathcal{I}} \left(U^i_{r,t} + W^i_{r,t} + E^i_{r,t} + M^i_{r,t} + D^i_{r,t}\right) \\ & \qquad + \sum_{r,t}^{i} D^i_{r,t}\left[\theta^{j,i}_{r,treat} - \theta^{j,i}_{r,untreat}\right] + N^i_{r,t} \qquad (3.40) \\ & = \sum_{j\in\mathcal{I}} \left(K^i_{r,t} + \sum_{h\in\mathcal{I},s\in\{1,2,\dots,t\}} v^i_{r,s} D^i_{r,t} \right) \theta^{j,i}_{r,untreat} + \sum_{r,t}^{i} D^i_{r,t}\left[\theta^{j,i}_{r,treat} - \theta^{j,i}_{r,untreat}\right] + N^i_{r,t} \\ & = K^i_{r,t+1} + \sum_{i\in\mathcal{I},s\in\{1,2,\dots,t\}} v^i_{r,s} D^i_{r,s} \end{aligned}$$

Where, $K^i_{r,t+1} = \sum_{j\in\mathcal{I}} K^i_{r,t}\theta^{j,i}_{r,untreat} + Nr,t+1^i$, $v^{j,i}_{r,t+1} = [\theta^{j,i}_{r,treat} - \theta^{j,i}_{r,untreat}]$ and $v^{j,i}_{r,t} = \sum_{h\in\mathcal{I},j\in\mathcal{I},s\in\{1,2,\dots,t\}} v^{h,j}_{r,s}\theta^{j,i}_{r,untreat}$. This shows that if (3.39) holds for t, it also holds for $t+1$. Therefore, by induction (3.39) is true. We next substitute (3.39) in the objective function of QMPPB to get:

$$\sum_{j\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i \left(K^i_{r,t} + \sum_{j\in\mathcal{I}} v^{j,i}_{r,s} D^j_{r,s} \right)$$

After simplifying,

$$K_0 + \sum_{j\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} \pi^i_{r,t} D^i_{r,t}$$

Where, $K_0 = \sum_{j\in\mathcal{I},r\in\mathcal{R},t\in\mathcal{M}_\tau,\tau\in[T]} q^i K^i_{r,t}$ and $\pi^i_{r,t} = \sum_{j\in\mathcal{I},s\in\{1,2,\dots,t\}} q^i v^{j,i}_{r,s}\square$

Proof of Proposition 7

Consider inequality (3.10) of QMPP:

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_o-\{13\},r\in\mathcal{R}} U^i_{r,t} \leq \min_k \{n_{k,P}A_{k,P}/y_{k,P}\} + \frac{1}{\tau_P}\ln(1-\alpha_P) \quad \forall t$$

This can be written as,

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R}} \alpha\beta S_{r,t} U^i_{r,t} + \sum_{i\in\mathcal{I}_o-\{13\},r\in\mathcal{R}} U^i_{r,t} \leq n_{k,P}A_{k,P}/y_{k,P} + \frac{1}{\tau_P}\ln(1-\alpha_P) \quad \forall t, k$$

Replacing $U^i_{r,t}$ with its lower bound, $\underline{U}^i_{r,t}$ and rearranging terms, we get,

$$\left(\frac{y_{k,P}}{A_{k,P}}\right)\left(\sum_{j\in\mathcal{I}_w,r\in\mathcal{R}}\alpha\beta S_{r,t}\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}_o-\{13\},r\in\mathcal{R}}\right)-\frac{y_{k,P}}{A_{k,P}\tau_P}\ln\left(1-\alpha_P\right)\le n_{k,P}\forall k$$

multiplying w_k and summing across k constraints,

$$\sum_{k\in\mathcal{W}}\left\{\left(\frac{y_{k,P}}{A_{k,P}}\right)\left(\sum_{j\in\mathcal{I}_w,r\in\mathcal{R}}\alpha\beta S_{r,t}\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}_o-\{13\},r\in\mathcal{R}}\right)-\frac{y_{k,P}}{A_{k,P}\tau_P}\ln\left(1-\alpha_P\right)\right\}$$
$$\le\sum_{k\in\mathcal{W}}w_k n_{k,P}$$

Similarly from (3.12) we get,

$$\sum_{i\in\mathcal{I}-\{13\},r\in\mathcal{R}}\left(M^i_{r,t}\phi^i_M+D^i_{r,t}\phi^i_D\right)=\min_k\left\{n_{k,S}A_{k,S}/y_{k,S}\right\}+\frac{1}{\tau_S}\ln\left(1-\alpha_S\right)\qquad\forall t$$

$$\sum_{k\in\mathcal{W}}\left\{\left(\frac{w_k y_{k,S}}{A_{k,S}}\right)\sum_{i\in\mathcal{I}-\{0\}.r\in\mathcal{R}}D^i_{r,t}\phi^i_D-\frac{w_k y_{k,S}}{A_{k,S}\tau_S}\ln(1-\alpha_S)\right\}\le\sum_{k\in\mathcal{W}}w_k n_{k,S}$$

Consider inequality (3.8)

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\alpha\beta S_{r,t}U^i_{r,t}+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^iUr,t^i$$
$$+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau,X\in\mathcal{X}}+\sum_{\ell\in\mathcal{L},k\in\mathcal{W},t\in\mathcal{M}\tau}n_{k,l}w_k\le B(\tau)\qquad\forall\tau$$

Replacing $U^i_{r,t}$ by $\underline{U}^i_{r,t}$ and dropping non-negative terms,

$$\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\alpha\beta S_{r,t}\underline{U}^i_{r,t}+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau}CS^i\underline{U}r,t^i+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau}C^i_U\underline{U}r,t^i$$
$$+\sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau}C^i_D D^i_{r,t}+\sum_{\ell\in\mathcal{L},k\in\mathcal{W},t\in\mathcal{M}\tau}n_{k,l}w_k\le B(\tau)\qquad\forall\tau$$

Substituting from above,

$$\begin{aligned}
&\sum_{i\in\mathcal{I}_w,r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i\alpha\beta S_{r,t}\underline{U}^i_{r,t} + \sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau} CS^i\underline{U}r,t^i + \sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau} C^i_U\underline{U}r,t^i \\
&\qquad + \sum_{i\in\mathcal{I}_o,r\in\mathcal{R},t\in\mathcal{M}_\tau} C^i_D D^i_{r,t} \\
&\sum_{k\in\mathcal{W}}\Bigg\{\left(\frac{y_{k,P}}{A_{k,P}}\right)\left(\sum_{j\in\mathcal{I}_w,r\in\mathcal{R}}\alpha\beta S_{r,t}\underline{U}^i_{r,t} + \sum_{i\in\mathcal{I}_o-\{13\},r\in\mathcal{R}}\right) \\
&\qquad - \frac{y_{k,P}}{A_{k,P}\tau_P}\ln\left(1-\alpha_P\right)\Bigg\} \\
&\sum_{k\in\mathcal{W}}\left\{\left(\frac{w_k y_{k,S}}{A_{k,S}}\right)\sum_{i\in\mathcal{I}-\{0\}.r\in\mathcal{R}} D^i_{r,t}\phi^i_D - \frac{w_k y_{k,S}}{A_{k,S}\tau_S}\ln(1-\alpha_S)\right\} \le B(\tau)
\end{aligned}$$

This simplifies to,

$$\sum_{r\in\mathcal{R},t\in\mathcal{M}_\tau}\sigma_{r,t}S_{r,t} \le B(\tau) - K_\tau - \sum_{i\in\mathcal{I}-\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}\rho^i D^i_{r,t} \tag{3.41}$$

This is the first inequality in the proposition with the associated definitions of K_τ, ρ^i, $\sigma_{r,t}$ and $\sigma_{r,t}$. Note that the total number of patients treated in each risk category, has to be less than the total number of patients screened and the total number of unscreened patients who get infected with OI. Thus:

$$\sum_{i\in\mathcal{I},t\in\mathcal{M}_\tau} \le \sum_{i\in\mathcal{I},t\in\mathcal{M}_\tau}\alpha\beta S_{r,t}\bar{U}^i_{r,t} + \sum_{i\in\mathcal{I},t\in\mathcal{M}_\tau}\bar{U}^i_{r,t}$$

The above inequality can be rearranged to get the second inequality in the proposition.

For stationary screening, setting $S_{r,t} = S_r\ \forall t$, we get

$$S_r \le \frac{B(\tau) - K_\tau}{\sum_{t\in\mathcal{M}_\tau}\sigma_{r,t}}$$

Search Algorithm for Stationary Screening

```
Start: $\Delta S$, $i \leftarrow 0, j \leftarrow 0, S_{hi} \leftarrow 0, S_{lo} \leftarrow 0$, $MAX \leftarrow 0$, $N \leftarrow 1/\Delta S$,
    $\bar{S}_{lo} = \min_{\tau} \frac{B(\tau) - K_t au}{\sum_{t \in \mathcal{M}_\tau} \sigma_{lo,\tau}}$, $\bar{S}_{hi} = \min_{\tau} \frac{B(\tau) - K_t au}{\sum_{t \in \mathcal{M}_\tau} \sigma_{hi,\tau}}$
while $i < N + 1$ and $S_{lo} < \bar{S}_{lo}$ do
    $S_{lo} \leftarrow S_{lo} + i\Delta S$
    $j \leftarrow 0$
    while $j < N + 1$ and $S_{hi} \leq \bar{S}_{hi}$ do
        $S_{hi} \leftarrow S_{hi} + j\Delta S$
        Evaluate $QMPPB(S_{hi}, S_{lo})$
        if $QMPPB(S_{hi}, S_{lo})$ is infeasible then
            End Do
        if $MAX < QMPPB(S_{hi}, S_{lo})$ then
            $MAX \leftarrow QMPPB(S_{hi}, S_{lo})$
            $S_{opt,hi} \leftarrow S_{hi}$
            $S_{opt,lo} \leftarrow S_{lo}$
        $j \leftarrow j + 1$
        End Do
    $i \leftarrow i + 1$
    End Do
```

Budget Imputation Algorithm

Procedure for choosing the number of partitions in the upper bound calculation

In determining the upper bound for the QMPP we need to choose parameter , the number of partitions on $U^i_{r,t}$. Note that as m increases, the value of the upper bound decreases (or becomes tighter) but its computation time becomes larger. Our procedure chooses by comparing this reduction of the bound value with its increase in computation time. To initialize this procedure, we start with $m = 1$ and record the value of the upper bound along with the time GUROBI takes to compute the bound. Next, we increment m by 1 and calculate the % reduction of the value of the bound and the % increase in computation time from the previous value of m. We then calculate the efficiency ratio defined as (% reduction in bound value)(% increase in computation time) and choose m corresponding to the highest ratio. We applied this procedure to our data for $m = 1$ to 7 as GUROBI was unable to solve upper bounds for $m > 7$. We found the best choice was at $m = 5$.

Estimation of system state costs C^i_X

C^i_X is composed of the following:

1. In Patient costs (CI^i): The average in-patient costs, (CI^i) per patient per month was collected from VHA data. This cost is incurred on all the patients at each system state. Thus, the in-patient cost is:

$$CI^i = \left(\alpha U^i_{r,t} + W^i_{r,t} + E^i_{r,t} + M^i_{r,t} + D^i_{r,t}\right)$$

2. Monitoring costs (CM^i): The monthly per-patient monitoring costs CM^i, is incurred on patients under monitoring $M^i_{r,t}$,as well as treatment $D^i_{r,t}$. This

```
Start: Set $S_{lo,t} \leftarrow 0$, $S_{hi,t} \leftarrow 1$, $\Delta B \leftarrow \$0.5mn$, $count \leftarrow 0$, $B \leftarrow 0$, $obj \leftarrow 0$,
    $\underline{\mathrm{B}} \leftarrow 0$, $MAXQ \leftarrow 0$, $\bar{B} \leftarrow 0$, $EXIT \leftarrow 0$
while $EXIT = 0$ do
    Evaluate $QMPPB(S_{r,t}, B)$
    if $QMPPB(S_{r,t}, B)$ is feasible then
        $B \leftarrow QMPPB(S_{r,t}, B)$
        $EXIT \leftarrow 1$
    else
        $count \leftarrow count + 1$
        $B \leftarrow B + \Delta B * count$
End Do
$EXIT \leftarrow 0$
$B \leftarrow \underline{B}$
while $EXIT = 0$ do
    Evaluate $QMPPB(S_{r,t}, B)$
    if $MAXQ > QMPPB(S_{r,t}, B)$ then
        $\bar{B} \leftarrow QMPPB(S_{r,t}, B)$
        $EXIT \leftarrow 0$
    else
        $count \leftarrow count + 1$
        $B \leftarrow B + \Delta B * count$
End Do
```

is the cost of one CD4 cell count and one HIV-1 RNA quantitation, per quarter. Anaya et al. (2012) provide the cost of CD4 cell count and RNA quantitation. The monitoring cost is:

$$CM^i = (M^i_{r,t} + D^i_{r,t})$$

3. Treatment costs (CT^i): The treatment cost per patient CT^i is the cost of pharmacy for patients undergoing treatment under HAART. The treatment cost is :

$$CT^i D^i_{r,t}$$

Outpatient overhead costs ($Coh^i_{X,M}$): The per patient overhead costs, $Coh^i_{X,M}$, was not directly available. Only the per-patient outpatient cost CO^i, was available from VA. This cost however, was inclusive of monitoring test costs and labor costs, which have already been accounted in the monitoring costs described above and in wages. Thus, in order to calculate outpatient overhead costs, we need to subtract the monitoring costs and the labor cost is:

$$Coh^i_{X,M} = CO^i - CM^i - L^i_X$$

Here,L^i_X is the out-patient labor utilization cost per patient at system state X. Let $y_{X,k}$ denote the labor time of staff of type k, required per patient visit at system state X. Further, let w_k denote the wage per time of staff type k and the ϕ^i_X the frequency of visits. These are them used to calculate the labor cost incurred per patient per month as

$$L^i_X = \phi^i_M(\sum_{k \in \mathcal{W}} y_{k,X} w_k)$$

Since outpatient overhead cost is incurred on all patients in the system, the total outpatient overhead cost for year τ would be given by:

$$\sum_{i \in \mathcal{I}-\{13\}, X \in \mathcal{W}} \left[Cok^i_X \mathbf{1}_{X=U} \alpha X^k_{r,t} + Cok^i_X \mathbf{1}_{X \neq U} X^k_{r,t}\right]$$

Collecting the costs, we get,

$$\sum_{i\in\mathcal{I}-\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau}\Bigg[(Coh_U^i\alpha+CI^i\alpha)U_{r,t}^i+(Coh_W^i+CI^i)W_{r,t}^i$$
$$+(Coh_E^i+CI^i)E_{r,t}^i+(Coh_M^i+CM^i+CI^i)M_{r,t}^i$$
$$+(Coh_D^i+CD^i+CI^i+CT^i)D_{r,t}^i\Bigg]$$

Collecting the terms in order to simplify the notation the total costs can be written as,

$$\sum_{i\in\mathcal{I}-\{13\},r\in\mathcal{R},t\in\mathcal{M}_\tau,X\in\mathcal{X}}\left[C_X^iX_{r,t}^i\right]$$

Where,

$$\begin{aligned}C_U^i&=\alpha(CI^i+Coh_U^i)\\C_W^i&=\alpha(CI^i+Coh_W^i)\\C_E^i&=\alpha(CI^i+Coh_E^i)\\C_M^i&=\alpha(CI^i+Coh_M^i+CM^i)\\C_D^i&=\alpha(CI^i+Coh_D^i+CD^i+CT^i)\end{aligned}$$

For brevity, we report refer the reader to the electronic companion of Deo et al.[4] Further Detailed breakdown are available upon request from the authors.

Computation of Transition Rates

As discussed in the paper, there are four processes which govern the transition from one health state to another: 1) HIV infection, 2) HIV infection progression (treated and untreated), 3) Opportunistic infection (OI), and 4) OI recovery.

[4] Available at `http://dx.doi.org/10.1287/opre.2015.1353`

The first process is the HIV infection process which governs the transition from health state 0 (uninfected) to health state 1 ($> 500cells/mm^3$). The monthly rate of transition under the HIV Infection process is denoted by $\theta^{0,1}_{r,untreat}$ where $\theta^{0,1}_{r,untreat} = incid_r/12$, where $incid_r$ is the annual incidence rate of risk category . We used the estimates provided by Paltiel et al. (2005) for the incidence rates ($incid_r$) This is shown provided in Deo et al.

The HIV progression process governs progression from one infected state to a higher infected state. The transition rate of this process varies depending on whether the patient is undergoing Highly Active Anti-Retroviral Treatment (HAART) or not. This transition rate from infected stage i to infected stage j for risk category r is given by $\theta^{ij}_{r,treat}$ and $\theta^{ij}_{r,untreat}$ for patients under HAART and not under treatment respectively. Mauskopf et al. (2005), calculate $p^{i,j}_{6-month}$ the six month transition probabilities from one health state to another without treatment. These 6 month transition probabilities are used to calculate monthly rate as $\theta^{ij}_{r,treat} = 1 - (1 - p^{i,j}_{6-month})/6$. Mauskopf et al. (2005) also provide relative risk of transition ($relrisk^{i,j}_{TR}$) between states in different treatment regimens (TRs), namely, First-line, Second-line, Salvage, and Optimized Background therapies. This relative risk is used to calculate the transition rates under each treatment regimen. The transition rate under treatment regimen TR is given as $\theta^{ij}_{r,TR} = \theta^{ij}_{r,untreat}(1 - relrisk_{TR})$. The overall transition rate under treatment is given by average of the transition rates under different treatment regimens or:

$$\theta^{ij}_{r,treat} = (\theta^{ij}_{r,first-line} + \theta^{ij}_{r,second-line} + \theta^{ij}_{r,salvage} + \theta^{ij}_{r,optimized})/4$$

The third process is the OI process that relates to patients infected with HIV who are susceptible to such infections. The rate with which they can be infected with these infections depends on the nature of the opportunistic infection and the current CD4 state of the patient. This transition rate is given by $\theta^{i,i+6}_{r,treat}$ and $\theta^{i,i+6}_{r,untreat}$ where $i \in \mathcal{I}_w$. Paltiel et al. (2005) provide the monthly risk of being infected with OI by CD4 stratum. For each CD4 category, we sum across the different OI to calculate the average risk of infection of OI.

Finally, the OI recovery process governs the recovery from such infection. The transition rates here are given by $\theta^{i,i+6}_{r,untreat}$. Kaplan et al. (2009) provide typical time required for recovery from each OI. The typical recovery times are converted to a weighted average recovery time using the relative risk of incurring that OI. This weighted average monthly recovery time is converted to the fraction or rate of patients recovering every month by $1 - e^{1.06} = 0.654$. Thus, the transition rate from any OI infected state to OI uninfected state of the same CD4 bracket $\theta^{i,i+6}_{r,untreat}$, is 0.654.

For transitions that require two processes to occur simultaneously such as transition between health states and transition to an OI status, we assume independence. Thus, the rates of the two processes occurring simultaneously are the product of the rates of the individual processes.

The transition rates are provided in the online companion of Deo et al.

Estimation of Quality of Life Utilities

The Quality of Life (QOL) utilities are drawn from two sources,Mauskopf et al. (2005) and Freedberg et al. (1998). Specifically, Mauskopf et al. (2005) provides 5 CD4 ranges, $\geq 500 cells/\mu L$, $350 - 499 cells/\mu L$, $200 - 349 cells/\mu L$, $100 - 199 cells/\mu L$ and $0 - 100 cells/\mu L$ and death. We further divide the range $0 - 100 cells/\mu L$ into two, $50 - 99 cells/\mu L$ and $0 - 49 cells/\mu L$ because the treatment and system costs for these two CD4 ranges were different (Schackman et al., 2006). These health states are numbered 1 through 6 and death. The QOL utilities for health states 1-4 was from Table 2 in Mauskopf et al. (2005). The QOL utilities for health states 5 and 6 were from Table 3.3 in Freedberg et al. (1998) By definition, the no infection state 0 has a QOL utility 1 and the death state 13 has a QOL utility 0.

Based on discussions with the physicians at the GLA station, we also incorporated health states with opportunistic infections by adding health states 7 through 12. As shown in Table 1, each of these states correspond to the same CD4 counts as in states 1 through 6 respectively, but have opportunistic infections. For example health state 7 (i.e., CD4 500 cells/L) corresponds to the CD4 count of health state 1, health state 8 with health state 2, and so on. The QOL utility for health states 7-12 were calculated in Freedberg et al. (1998) Here, we considered the health related quality adjustment scores for the opportunistic infections by listed pathogen types (such as Pneumocystis Carini, through other AIDS diagnoses). Ideally, one would have had to introduce additional sub health states for each opportunistic infection within a CD4 count range. However, the physicians felt that it would be impractical to do since patients typically had more than one opportunistic infection, it was often not easy to diagnose the pathogens and decide which one was most dominant. Further, the range of the scores across these opportunistic infections was relatively narrow (i.e., 0.56 to 0.65). Therefore, it was considered reasonable to calculate the quality utility for health states 7 through 12 by averaging the quality scores across these opportunistic infections.

CHAPTER 4

Future Research

In this dissertation I consider two resource planning problems. The two problems differ in context and methodology. However, they touch upon some of the characteristics making healthcare resource planning a complex issue, in particular,

- Large scale of operations at major hospitals
- Interplay between clinical decision making and operational decision making
- Uncertainty in resource utilization
- Expensive, specialized resources

In both the VHA GLA and UCLA Ronald Reagan Medical Center, the critical resources were the human resources; physicians, nurses, counselors, technicians and anesthesiologists. In fact almost 50% of expenditure of aggregate healthcare expenditure are salaries (Glied et al., 2015).

Particular care must taken when creating resource scheduling and planning systems involving human resources. Scheduling and planning systems must be designed not to tightly control workers but to inform planners for decision making. Recent research has shown that algorithm driven solutions are more acceptable when the scheduling systems incorporate the needs of the workers as well (Bernstein et al., 2014; Dietvorst et al., 2015).

One way to incorporate employee satisfaction is to model inconvenience costs of plans. However, these costs are not explicitly known. We look at the monthly planning for anesthesiologists at the UCLA Ronald Reagan Medical Center as a problem where the planning requires such inconvenience costs as input. As future research we propose to use structural estimation approaches to infer these inconvenience costs from historical decisions of the planners.

The Anesthesiologist Planning Problem

Introduction Planning for anesthesiologists at UCLA Ronald Reagan Medical Center is performed centrally for its 23 Operating Rooms, 8 interventional procedure rooms, 6 catheterization labs and 3 medical procedure units. Wide variety of procedures are conducted in a day and each procedure length is unpredictable, therefore the demand for anesthesia service on any given day has considerable variability. Additionally, due to the research and teaching responsibilities and vacation schedules of anesthesiologists, the availability of any particular anesthesiologists is known only a few weeks in advance. Due to these factors, the staff planning for anesthesiologists is conducted under uncertainty in the availability of anesthesiologists and the demand for anesthesia services. To enable management to react to a changing environment, staff planning is done in a several stages consisting of annual, monthly and daily decisions. particular methodology for staff planning followed at the monthly decision making is called the *Q-Call* system. The Q-Call system is a way for staff planners to react to uncertainty in demand. Anesthesiologists available on a Q-Call consideration list for a particular day can be brought in to work on that day by informing them the day before. These anesthesiologists are paid an additional $1000 per day over their regular salary. The anesthesiologists on the Q-Call consideration list thus act as a reactive capacity who can be brought in at an additional cost once the number of elective procedures to be performed the next day is known. Similar systems are used for anesthesiologists and nursing staff at most major hospitals.

The decision time-line of the planning is as follows:

1. On a longer term planning basis the decision to hire additional anesthesiologists of particular specialties is made. This decision is made based on a forecast of the annual demand and the strategic focus of the hospital towards particular specialties.

2. On the 20th of each month, the availability of anesthesiologists for the next month is known based on their teaching and vacation schedule. Based on this information the daily plan for anesthesiologists for the next month is prepared. This plan consists of creating two lists of anesthesiologists for each day of the next month: those to be available on regular duty and those on the Q-Call consideration list.

3. The day before the surgery the total number of elective procedures to be performed is known. Based on this information a certain number of anesthesiologist from the Q-Call consideration list are informed that they would be working the next day .On the day of the surgery, the actual demand for anesthesia services is realized based on the actual duration of procedures and any emergency cases. Thus, at the end of the day the total overtime hours or idle hours are realized.

Motivation For the monthly tactical planning, the staff planners minimize the sum of expected Q-Call, overtime and idle time costs. This decision is complex due to two reasons, first, the high degree of uncertainty in demand for anesthesia services and second, some cost components are not explicitly known. The costs of overtime and the cost of calling an anesthesiologist are known. However, the cost of idle time and the cost of keeping an anesthesiologist on the consideration list but not calling them are not explicitly known. While the hospital does not make direct payments for these, they are an important component of decision making as high idle time and not utilizing anesthesiologist from the Q-Call consideration list impacts employee morale and leads to dissatisfied employees and thus would have longer term implications. Interviews with planners revealed that while they did not have a numerical value for these costs, they implicitly take these into account while making their monthly plan. The management thus balances these costs while creating the monthly plan. An optimization model that does not incorporate both explicit and implicit costs would be incomplete.

Problem Statement To use historical data in a structural estimation framework to infer the implicit costs of Q-call and idle time. To draw managerial insights from these inferred costs and to use these inferred costs to develop an optimization model for the monthly staff planning. The structural estimation of these implicit costs also has the potential to aid decision making in the hiring of anesthesiologists.

Bibliography

Jean Abadie and J Carpentier. Generalization of the wolfe reduced gradient method to the case of nonlinear constraints. *Optimization*, 37:47, 1969.

AMA. http://www.ama-assn.org/ama/pub/news/news/2012-09-17-cpt-code-changes-2013.page, 2013.

Henry D Anaya, Kee Chan, Uday Karmarkar, Steven M Asch, and Matthew Bidwell Goetz. Budget impact analysis of hiv testing in the va healthcare system. *Value in Health*, 15(8):1022–1028, 2012.

Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

David W Bates and Atul A Gawande. Improving safety with information technology. *New England journal of medicine*, 348(25):2526–2534, 2003.

Sakine Batun, Brian T Denton, Todd R Huschka, and Andrew J Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS J. Comput.*, 23(2):220–237, 2011.

Jeroen Beliën and Erik Demeulemeester. A branch-and-price approach for integrating nurse and surgery scheduling. *Eur. J. Oper. Res.*, 189(3):652–668, 2008.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Ethan Bernstein, Saravanan Kesavan, and Staats Bradley. How to Manage Scheduling Software Fairly. `https://hbr.org/2014/09/how-to-manage-scheduling-software-fairly`, 2014.

Dimitris Bertsimas and Melvyn Sim. The price of robustness. *Oper. Res.*, 52(1): 35–53, 2004.

Dimitris Bertsimas and Aurélie Thiele. A robust optimization approach to inventory theory. *Oper. Res.*, 54(1):150–168, 2006.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *arXiv preprint arXiv:1401.0212*, 2013.

John R Birge and Francois Louveaux. *Introduction to stochastic programming.* Springer Science & Business Media, 2011.

John R Birge and Francois V Louveaux. A multicut algorithm for two-stage stochastic linear programs. *Eur. J. Oper. Res.*, 34(3):384–392, 1988.

David Bishai, Arantxa Colchero, and David T Durack. The cost effectiveness of antiretroviral treatment strategies in resource-limited settings. *Aids*, 21(10): 1333–1340, 2007.

John T Blake and Michael W Carter. Surgical process scheduling: a structured review. *Journal of the Society for Health Systems*, 5(3):17–30, 1996.

S Blount, A Galambosi, and S Yakowitz. Nonlinear and dynamic programming for epidemic intervention. *Applied Mathematics and Computation*, 86(2):123–136, 1997.

Margaret L Brandeau, Gregory S Zaric, and Anke Richter. Resource allocation for control of infectious diseases in multiple independent populations: beyond cost-effectiveness analysis. *Journal of health economics*, 22(4):575–598, 2003.

Jennifer Bresnick. Healthcare big data analytics driving billions in market growth. http://healthitanalytics.com/news/healthcare-big-data-analytics-driving-billions-in-market-growth, 2015. Accessed: 2010-09-06.

Brecht Cardoen, Erik Demeulemeester, and Jeroen Beliën. Operating room planning and scheduling: A literature review. *Eur. J. Oper. Res.*, 201(3):921–932, 2010a.

Brecht Cardoen, Erik Demeulemeester, and Jessie Van der Hoeven. On the use of planning models in the operating theatre: results of a survey in flanders. *The International journal of health planning and management*, 25(4):400–414, 2010b.

Claus C Carøe and Rüdiger Schultz. Dual decomposition in stochastic integer programming. *Operations Research Letters*, 24(1):37–45, 1999.

Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

Robert R Cima, Michael J Brown, Hebl, et al. Use of lean and six sigma methodology to improve operating room efficiency in a high-volume tertiary-care academic medical center. *Journal of the American College of Surgeons*, 213(1): 83–92, 2011.

Brian Denton and Diwakar Gupta. A sequential bounding approach for optimal appointment scheduling. *IIE Transactions*, 35(11):1003–1016, 2003.

Brian Denton, James Viapiano, and Andrea Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science*, 10(1):13–24, 2007.

Brian T Denton, Andrew J Miller, Hari J Balasubramanian, and Todd R Huschka. Optimal allocation of surgery blocks to operating rooms under uncertainty. *Oper. Res.*, 58(4-part-1):802–816, 2010.

Sarang Deo, Seyed Iravani, Tingting Jiang, Karen Smilowitz, and Stephen Samuelson. Improving health outcomes through better capacity allocation in a community-based chronic care model. *Oper. Res.*, 61(6):1277–1294, 2013.

Franklin Dexter and Rodney D Traub. How to schedule elective surgical cases into specific operating rooms to maximize the efficiency of use of operating room time. *Anesthesia & Analgesia*, 94(4):933–942, 2002.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use algorithms if they can (even slightly) modify them. *Available at SSRN 2616787*, 2015.

E Dolan, Robert Fourer, Jorge J Moré, Munson Munson, et al. The neos server for optimization: Version 4 and beyond. *Preprint ANL/MCS-TM-253, Mathematics and Computer Science Division, Argonne National Laboratory*, 2002.

Paul Dolan, Rebecca Shaw, Aki Tsuchiya, and Alan Williams. Qaly maximisation and people's preferences: a methodological review of the literature. *Health economics*, 14(2):197–208, 2005.

Arne Drud. Conopt: A grg code for large sparse dynamic nonlinear optimization problems. *Mathematical Programming*, 31(2):153–191, 1985.

Kenneth A Freedberg, Julie A Scharfstein, George R Seage III, Elena Losina, Milton C Weinstein, Donald E Craven, and A David Paltiel. The cost-effectiveness of preventing aids-related opportunistic infections. *Jama*, 279(2):130–136, 1998.

Neel R Gandhi, Melissa Skanderson, Kirsha S Gordon, John Concato, and Amy C Justice. Delayed presentation for human immunodeficiency virus (hiv) care among veterans: a problem of access or screening? *Medical care*, 45(11):1105, 2007.

Atul Gawande. *Complications: A surgeon's notes on an imperfect science*. Profile Books, 2010.

Ramsis F Ghaly. Do neurosurgeons need neuroanesthesiologists? should every neurosurgical case be done by a neuroanesthesiologist? *Surgical Neurology International*, 5, 2014.

Risha Gidwani, Matthew Bidwell Goetz, Gerald Kominski, Steven Asch, Kristin Mattocks, Jeffrey H Samet, Amy Justice, Neel Gandhi, and Jack Needleman. A budget impact analysis of rapid human immunodeficiency virus screening in veterans administration emergency departments. *The Journal of emergency medicine*, 42(6):719–726, 2012.

Sherry A Glied, Stephanie Ma, and Ivanna Pearlstein. Understanding pay differentials among health professionals, nonprofessionals, and their counterparts in other sectors. *Health Affairs*, 34(6):929–935, 2015.

Matthew B Goetz, Candice Bowman, Tuyen Hoang, Henry Anaya, Teresa Osborn, Allen L Gifford, and Steven M Asch. Implementing and evaluating a regional strategy to improve testing rates in va patients at risk for hiv, utilizing the queri process as a guiding framework: Queri series. *Implement Sci*, 3(1):16, 2008a.

Matthew Bidwell Goetz and David Rimland. Effect of expanded hiv testing programs on the status of newly diagnosed hiv-infected patients in two veterans health administration facilities: 1999–2009. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 57(2):e23–e25, 2011.

Matthew Bidwell Goetz, Tuyen Hoang, Candice Bowman, Herschel Knapp, Barbara Rossman, Robert Smith, Henry Anaya, Teresa Osborn, Allen L Gifford, Steven M Asch, et al. A system-wide intervention to improve hiv testing in the veterans health administration. *Journal of general internal medicine*, 23 (8):1200–1207, 2008b.

Matthew Bidwell Goetz, Tuyen Hoang, Herschel Knapp, S Randal Henry, Henry Anaya, Ann F Chou, Allen L Gifford, and Steven M Asch. Exportability of an intervention to increase hiv testing in the veterans health administration. *The Joint Commission Journal on Quality and Patient Safety*, 37(12):553–559, 2011.

Matthew Bidwell Goetz, Tuyen Hoang, Herschel Knapp, Jane Burgess, Michael D Fletcher, Allen L Gifford, Steven M Asch, QUERI-HIV/Hepatitis Program, et al. Central implementation strategies outperform local ones in improving hiv testing in veterans healthcare administration facilities. *Journal of general internal medicine*, 28(10):1311–1317, 2013.

Linda V Green. O. M. forum-The vital role of operations analysis in improving healthcare delivery. *Manufacturing Service Oper. Management*, 14(4):488–494, 2012.

Linda V Green and Sergei Savin. Reducing delays for medical appointments: A queueing approach. *Oper. Res.*, 56(6):1526–1538, 2008.

Peter Groves, Basel Kayyali, David Knott, and Steve Van Kuiken. The 'big data' revolution in healthcare. *McKinsey Quarterly*, 2013.

Francesca Guerriero and Rosita Guido. Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, 14(1): 89–114, 2011.

Diwakar Gupta. Surgical suites' operations management. *Production and Operations Management*, 16(6):689–700, 2007.

Inc. Gurobi Optimization. Gurobi optimizer reference manual, http://www.gurobi.com, 2015.

Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and Ronald L Tatham. *Multivariate Data Analysis*, volume 6. Pearson Prentice Hall Upper Saddle River, NJ, 2006.

Alex B Haynes, Thomas G Weiser, William R Berry, Lipsitz, et al. A surgical safety checklist to reduce morbidity and mortality in a global population. *New England Journal of Medicine*, 360(5):491–499, 2009.

Biyu He, Franklin Dexter, Alex Macario, and Stefanos Zenios. The timing of staffing decisions in hospital operating rooms: incorporating workload heterogeneity into the newsvendor problem. *Manufacturing Service Oper. Management*, 14(1):99–114, 2012.

Narges Hosseini, M Susan Hallbeck, Christopher J Jankowski, Jeanne M Huddleston, Amrit Kanwar, and Kalyan S Pasupathy. Effect of obesity and clinical factors on pre-incision time: Study of operating room workflow. In *AMIA Annual Symposium Proceedings*, volume 2014, page 691. American Medical Informatics Association, 2014.

William C Jordan and Stephen C Graves. Principles on the benefits of manufacturing process flexibility. *Management Sci.*, 41(4):577–594, 1995.

Jonathan E Kaplan, Constance Benson, King K Holmes, John T Brooks, Alice Pau, H Masur, Centers for Disease Control, Prevention (CDC), HIV Medicine Association of the Infectious Diseases Society of America, et al. Guidelines for prevention and treatment of opportunistic infections in hiv-infected adults and adolescents. *MMWR Recomm Rep*, 58(RR-4):1–207, 2009.

Enis Kayis, Haiyan Wang, Meghna Patel, Tere Gonzalez, Jain, et al. Improving prediction of surgery duration using operational and temporal factors. In *AMIA Annual Symposium Proceedings*, volume 2012, page 456. American Medical Informatics Association, 2012.

James E Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial & Applied Mathematics*, 8(4):703–712, 1960.

Leonard Kleinrock. *Theory, volume 1, Queueing systems.* Wiley-interscience, 1975.

Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Roger Koenker. *Quantile Regression.* Number 38. Cambridge university press, 2005.

Roger Koenker. Quantreg: Quantile regression. *R package version*, 5, 2013.

Panagiotis Kougias, Vikram Tiwari, Sonia Orcutt, Amber Chen, George Pisimisis, Neal R Barshes, Carlos F Bechara, and David H Berger. Derivation and out-of-sample validation of a modeling system to predict length of surgery. *The American Journal of Surgery*, 204(5):563–568, 2012.

Beste Kucukyazici, Vedat Verter, and Nancy E Mayo. An analytical framework for designing community-based care for chronic diseases. *Production and Operations Management*, 20(3):474–488, 2011.

Daniel M Laskin, A Omar Abubaker, and Robert A Strauss. Accuracy of predicting the duration of a surgical operation. *Journal of Oral and Maxillofacial Surgery*, 71(2):446–447, 2013.

Elisa F Long, Margaret L Brandeau, and Douglas K Owens. The cost-effectiveness and population outcomes of expanded hiv screening and antiretroviral treatment in the united states. *Annals of internal medicine*, 153(12):778–789, 2010.

Alex Macario. What does one minute of operating room time cost? *Journal of Clinical Anesthesia*, 22(4):233–236, 2010.

Ho-Yin Mak, Ying Rong, and Jiawei Zhang. Appointment scheduling with limited distributional information. *Management Sci.*, 2014a.

Ho-Yin Mak, Ying Rong, and Jiawei Zhang. Sequencing appointments for service systems using inventory approximations. *Manufacturing Service Oper. Management*, 16(2):251–262, 2014b.

Camilo Mancilla and Robert Storer. A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, 44(8): 655–670, 2012.

A Märkert and R Gollmer. User's guide to ddsip–AC package for the dual decomposition of two-stage stochastic programs with mixed-integer recourse, 2008.

Inês Marques, M Eugénia Captivo, and Margarida Vaz Pato. Scheduling elective surgeries in a portuguese hospital using a genetic heuristic. *Operations Research for Health Care*, 3(2):59–72, 2014.

Erika G Martin, A David Paltiel, Rochelle P Walensky, and Bruce R Schackman. Expanded hiv screening in the united states: what will it cost government discretionary and entitlement programs? a budget impact analysis. *Value in Health*, 13(8):893–902, 2010.

Richard Kipp Martin. *Large Scale Linear and Integer Optimization: A Unified Approach*. Springer Science & Business Media, 1999.

Josephine Mauskopf, Mari Kitahata, Teresa Kauf, Anke Richter, and Jerry Tolson. Hiv antiretroviral treatment: early versus later. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 39(5):562–569, 2005.

Garth P McCormick. Computability of global solutions to factorable nonconvex programs: Part iconvex underestimating problems. *Mathematical programming*, 10(1):147–175, 1976.

R McNicol. Paediatric anaesthesia–who should do it? the view from the specialist hospital. *Anaesthesia*, 52(6):513–515, 1997.

Nadine Meskens, David Duvivier, and Arnauld Hanset. Multi-objective operating room scheduling considering desiderata of the surgical team. *Decision Support Systems*, 55(2):650–659, 2013.

Daiki Min and Yuehwern Yih. Scheduling elective surgery under uncertainty and downstream capacity constraints. *Eur. J. Oper. Res.*, 206(3):642–652, 2010.

Seema U Nayak, Meredith L Welch, and Virginia L Kan. Greater hiv testing after veterans health administration policy change: the experience from a va medical center in a high hiv prevalence area. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 60(2):165–168, 2012.

Duncan Neuhauser, Lloyd Provost, and Bo Bergman. The meaning of variation to healthcare managers, clinical and health-services researchers, and individual patients. *BMJ quality & safety*, 20(Suppl 1):i36–i40, 2011.

Marcelo Olivares, Christian Terwiesch, and Lydia Cassorla. Structural estimation of the newsvendor model: an application to reserving operating room time. *Management Sci.*, 54(1):41–55, 2008.

Fredrick K Orkin, Sandra L Mcginnis, Gaetano J Forte, Peterson, et al. United States anesthesiologists over 50: retirement decision making and workforce implications. *Survey of Anesthesiology*, 57(2):101–102, 2013.

Douglas K Owens, Vandana Sundaram, Laura C Lazzeroni, Lena R Douglass, Gillian D Sanders, Kathie Taylor, Ronald VanGroningen, Vera M Shadle, Valerie C McWhorter, Teodora Agoncillo, et al. Prevalence of hiv infection among inpatients and outpatients in department of veterans affairs health care systems: implications for screening programs for hiv. *American journal of public health*, 97(12):2173–2178, 2007.

Frank J Palella, Maria Deloria-Knoll, Joan S Chmiel, Anne C Moorman, Kathleen C Wood, Alan E Greenberg, and Scott D Holmberg. Survival benefit of initiating antiretroviral therapy in hiv-infected persons in different cd4+ cell strata. *Annals of internal medicine*, 138(8):620–626, 2003.

A David Paltiel, Milton C Weinstein, April D Kimmel, George R Seage III, Elena Losina, Hong Zhang, Kenneth A Freedberg, and Rochelle P Walensky. Expanded screening for hiv in the united statesan analysis of cost-effectiveness. *New England Journal of Medicine*, 352(6):586–595, 2005.

Manuel Pardo. The development of education in anesthesia in the United States. In J. Fagerberg, D.C. Mowery, and R.R. Nelson, editors, *The Wondrous Story of Anesthesia*. Springer, 2014.

Laura A Petersen, Tracy H Urech, Margaret M Byrne, and Kenneth Pietz. Do financial incentives in a globally budgeted healthcare payment system produce changes in the way patients are categorized? a five-year study. *American Journal of Managed Care*, 13(9):513, 2007.

Marion S Rauner and Margaret L Brandeau. Aids policy modeling for the 21st century: an overview of key issues. *Health Care Management Science*, 4(3): 165–180, 2001.

Mark S Roberts, Kimberly A Nucifora, and R Scott Braithwaite. Using mechanistic models to simulate comparative effectiveness trials of therapy and to estimate long-term outcomes in hiv care. *Medical care*, 48(6):S90–S95, 2010.

http://newsroom.ucla.edu/releases/ucla-health-system-s-hospitals-ranked-among-nation-s-best-in-u-s-news-annual-survey Roxanne Moster, 2014.

Andrzej P Ruszczyński. *Nonlinear Optimization*, volume 13. Princeton university press, 2006.

N. V. Sahinidis. *BARON 14.3.1: Global Optimization of Mixed-Integer Nonlinear Programs,* User's Manual, 2014.

Andrew J Schaefer, Matthew D Bailey, Steven M Shechter, and Mark S Roberts. Modeling medical treatment using markov decision processes. In *Operations research and health care*, pages 593–612. Springer, 2005.

Steven M Shechter, Matthew D Bailey, Andrew J Schaefer, and Mark S Roberts. The optimal time to initiate hiv therapy under ordered health states. *Oper. Res.*, 56(1):20–33, 2008.

Abigail Silva, Nancy R Glick, Sheryl B Lyss, Angela B Hutchinson, Thomas L Gift, Lisa N Pealer, Dawn Broussard, and Steven Whitman. Implementing an hiv and sexually transmitted disease screening program in an emergency department. *Annals of emergency medicine*, 49(5):564–572, 2007.

David P Strum, Jerrold H May, and Luis G Vargas. Modeling the uncertainty of surgical procedure timescomparison of log-normal and normal models. *The Journal of the American Society of Anesthesiologists*, 92(4):1160–1167, 2000.

Aurélie Thiele, Tara Terry, and Marina Epelman. Robust linear optimization with recourse. *Rapport Technique*, pages 4–37, 2009.

Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014.

F.L. van Rossum, G Drake. Python reference manual, pythonlabs, virginia, USA, http://www.python.org, 2001.

Jagadisan Viswanathan and Ignacio E Grossmann. A combined penalty function and outer-approximation method for minlp optimization. *Computers & Chemical Engineering*, 14(7):769–782, 1990.

Danan Suryo Wicaksono and IA Karimi. Piecewise milp under-and overestimators for global optimization of bilinear programs. *AIChE Journal*, 54(4):991–1008, 2008.

Gregory S Zaric and Margaret L Brandeau. Resource allocation for epidemic control over short time horizons. *Mathematical Biosciences*, 171(1):33–58, 2001.

Gregory S Zaric, Margaret L Brandeau, and Paul G Barnett. Methadone maintenance and hiv prevention: a cost-effectiveness analysis. *Management Sci.*, 46 (8):1013–1031, 2000.