IMPROVING HEALTH CARE DELIVERY THROUGH MULTI-OBJECTIVE RESOURCE ALLOCATION

A Dissertation
Presented to
The Academic Faculty

by

Jacqueline A. Griffin

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2012

IMPROVING HEALTH CARE DELIVERY THROUGH MULTI-OBJECTIVE RESOURCE ALLOCATION

Approved by:

Professor Pinar Keskinocak, Advisor
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Dave Goldsman
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Martin Savelsbergh
School of Mathematical and Physical Sciences
University of Newcastle

Professor Turgay Ayer
School of Industrial and Systems Engineering
Georgia Institute of Technology

Professor Paul Griffin
Department of Industrial and Manufacturing Engineering
Penn State University

Date Approved: August 13, 2012

DEDICATION

In memory of

Frank Newman (1921-2011) and Alice Griffin (1925-2012)

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Pinar Keskinocak for her support and guidance during the process of completing this dissertation and degree. I would also like to acknowledge Dr. Martin Savelsbergh and Dr. Paul Griffin who served as advisors and mentors during my first years at Georgia Tech. Finally, I would like to thank the other members of my thesis committee, Dr. Dave Goldsman and Dr. Turgay Ayer, for their valuable feedback and advice.

While there are many Ph.D. students to thank, I would like to specifically acknowledge three. I have been lucky to have had many opportunities to collaborate with Hannah Smalley during our time at Georgia Tech. In addition to being a supportive friend, I have learned a lot from her through our work together. I have had wonderful officemates while at Georgia Tech including Antonio Carbajal. I would like to especially thank my officemate of six years, Kate Lindsey, who has provided me with friendship, support, encouragement, advice, and laughter as needed through the many ups and downs that have accompanied the completion of this degree and dissertation.

I am thankful for the many family members that have supported me through this process, including siblings, cousins, aunts, uncles, and grandparents. Specifically, I would like to thank my parents, Maureen Setteducato and Carl Griffin, who fostered my interest in mathematics from an early age and encouraged me in my goal of pursuing this degree.

Finally, I would like to thank Paul Kerl and Stout Griffin for their unending support. My dog, Stout, has sat by my side for much of the writing of this thesis, providing encouragement through his companionship and constantly wagging tail.

Most importantly, I would like to thank my boyfriend Paul for his love, patience, support, and help during this process. In addition to his daily words of encouragement he has dedicated his time to reading multiple iterations of my thesis and provided me with excellent feedback that has led to the betterment of my research and this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This dissertation addresses resource allocation problems that occur in both public and private health care settings with the objective of characterizing the tradeoffs that occur when simultaneously incorporating multiple objectives and developing methods to address these tradeoffs. We examine three resource allocation problems (i) strategic allocation of financial resources and limited staffing capacity for the mobile delivery of health care within African countries, (ii) real-time allocation of hospital beds to internal patient requests, and (iii) development of patient redirection policies in response to limited bed availability in units within a system of hospitals. For each problem we define models, each with a different methodology, and utilize the models to develop allocation strategies that account for multiple competing objectives and examine the performance of the strategies with computational studies.

In Chapter 2, we model African health care delivery systems utilizing a mixed-integer program (MIP) which accounts for financial and personnel constraints as well as infrastructure quality. We characterize tradeoffs in effectiveness, efficiency, and equity resulting from four allocation strategies with computational experiments representing the variety of spatial patterns that occur throughout the continent. The main contributions include (i) the development of a model that incorporates spatial and infrastructure characteristics and allows for a study of equity in the delivery of care, rather than access to care, and (ii) the characterization of tradeoffs in the three objectives under a variety of settings.

In Chapter 3, we model the real-time assignment of bed requests to available

beds as a queueing system and a Markov decision process (MDP). Through the development of bed assignment algorithms and simulation experiments, we illustrate the value of implementing strategic bed assignment practices which balance the bed management objectives of timeliness and appropriateness of assignments. The main contributions of this section include (i) the development of new bed assignment algorithms which use stochastic optimization techniques and outperform algorithms which mimic processes currently used in practice and (ii) the definition of a model and methods for the control of a large complex system that includes flexible units, multiple patient types, and type-dependent routing.

In Chapter 4, we model the impact of a patient redirection policy in a hospital unit as a Markov chain. Assuming preferences for patient redirection are aligned with costs, we examine the impact of incremental changes to redirection policies on the probability of the unit being completely occupied, the long-run average utilization, and the long-run average cost of redirection. The main contributions of this chapter include (i) the introduction of a model of patient redirection with multiple patient thresholds and patient preference constraints and (ii) the definition of necessary conditions for an optimal patient redirection policy that minimizes the average cost of redirection.

CHAPTER I

INTRODUCTION

A common problem addressed through operations research and mathematical modeling is the strategic allocation of limited resources to a set of jobs or activities. Decisions regarding the allocation of scarce resources must account both for the unique characteristics of the resources and tasks and the system objectives. Resource allocation problems arise in a variety of settings including financial portfolio management, project management, and manufacturing scheduling. Some resource allocation problems focus on strategic decision making, over a long horizon, while others focus on daily or hourly operational decisions. Additionally, resource allocation problems can be static or dynamic.

The management of most health care systems involves difficult resource allocation decisions which arise due to the imbalance between the supply of health services and patient needs. This imbalance is particularly apparent in the delivery of public health services in developing countries in which the level of global aid provided for health care is significantly outweighed by the resource requirements [43]. Similarly, an imbalance between the number of patients seeking care and the available capacity to treat patients occurs in hospitals throughout the United States, resulting in overcrowded hospitals and emergency departments [81].

Due to the delivery of life-altering services and the significant impact from poor decision making, resource allocation problems in health care systems often require a unique approach accounting for the unique system characteristics. Quality of health care delivery, which is one of the key factors in decision making, requires simultaneous consideration of a variety of objectives. The Institute of Medicine has identified six

aims for improving quality in health care delivery: effectiveness, efficiency, equity, timeliness, safety, and patient-centeredness [40]. The simultaneous consideration of these aims and their (often competing) interactions is necessary for informed decision making regarding the operation of health care systems and allocation of resources. As a result, improving quality within the confines of limited resources requires a multi-objective approach.

This dissertation focuses on the allocation of scarce resources in three health care system settings with respect to multiple objectives. In the first setting, we examine strategic decision making for the allocation of financial resources and limited staffing capacity in the delivery of health care services within African countries, accounting for the unique challenges that exist in this environment. In the second setting, we examine the real-time allocation of hospital beds to patient requests within a hospital, accounting for appropriateness and timeliness in patient care. Lastly, we examine the development of patient redirection policies within a system of hospitals, again allocating hospital beds to patient requests and accounting for a variety of system performance metrics. In all three settings, we utilize models to develop allocation strategies that account for multiple competing objectives and test the performance of the strategies through simulation and computational experiments.

1.1 African Health Care Delivery

In the first setting (Chapter 2), we model the decisions faced by managers within non-governmental organizations (NGO) or ministries of health (MoH) when structuring supply chains for delivery of health care services in developing countries. These decisions include the opening or use of specific health centers, allocation of funds and health supplies to the open health centers, and deployment of health center staff for administration within the communities.

Health care delivery in resource-constrained regions is made difficult by unreliable

transportation systems, the relative locations of health centers and settlements, and diverse geography [78, 85]. Therefore, the resource allocation model and policies we develop account for the logistical difficulties resulting from the distribution of populations and health centers as well as the structure and quality of the transportation infrastructure.

1.1.1 Objectives

Much of the difficulty in allocation of scare resources arises from the many competing objectives that must be considered. For example, goals include maximizing the health impact, while minimizing the resources that are utilized, and ensuring that allocation is fair and equitable. These objectives of effectiveness, efficiency, and equity are key elements in the delivery of quality health care in Africa [8, 85].

The logistical difficulties discussed above directly impact the effectiveness, efficiency, and equity of an allocation strategy. For example, due to the spatial characteristics of a region, it may be least expensive to the "health system" to distribute health services in densely-populated areas, which leads to limited access to care for communities in sparsely populated regions. Increasing equity by delivering health care to those in remote areas results in sacrifices in efficiency, with a significant increase in the average cost per patient treated [24]. Due to this increase in average patient cost, limited funds treat fewer patients thereby decreasing the overall effectiveness of the allocation. Therefore, a multi-objective approach that considers these tradeoffs is needed.

1.1.2 Methodology

To examine tradeoffs in effectiveness, efficiency, and equity we develop a mixed-integer program (MIP) to model an African health care system and the corresponding decisions for allocation of scarce resources. Unlike previous health allocation models [23], we incorporate objectives in a hierarchical fashion, identifying a solution with

the use of multiple objectives. We develop and examine four allocation strategies with computational experiments. Each strategy pertains to the order in which the objectives are used in a hierarchical optimization model.

1.1.3 Contribution

Rather than modeling one specific region in Africa, as many other studies have done, we create data instances reflecting a variety of settlement patterns often seen in Africa. By examining the strategies in multiple instances, we provide a comprehensive study of the impact of spatial characteristics and allocation strategies.

By comparing the allocation policies, corresponding to specific instances and strategies, we examine the interdependencies between the three objectives including the tradeoffs in equity and health improvement. Additionally we examine how the resource restrictions, budgetary and otherwise, impact the performance of the allocation strategies.

1.2 Dynamic Bed Assignment

In the remainder of the thesis, we address operational decision making in hospital systems. In Chapter 3 we address real-time bed assignment decisions faced by hospital bed managers. Bed management refers to the task of balancing supply of scarce hospital resources, primarily beds, and demand by way of patient admission and transfer requests [67].

In bed management, the bed assignment for a particular patient is linked to the nature of his or her condition. This allows for centralization of expertise by the nurses in a particular wing or area of the hospital, improved patient outcomes and safety, and more efficient rounding by physicians. While a patient's ideal bed assignment may be known, when a hospital is highly capacitated, a bed in the ideal unit may not be available and instead the patient may need to be placed in an overflow, or "off-service," unit with a potential detriment to effectiveness. An alternate option

is to postpone a patient's transfer, by continuing to "board" in the current unit, especially if none of the available beds meet the patient's medical needs. Despite the fact that a patient is not assigned to the ideal unit, both the choices to "board" in the current unit or assign a patient to an "overflow" unit correspond to decisions for the allocation of limited hospital resources.

1.2.1 Objectives

A decision to postpone an assignment, through "boarding," causes increased waiting by the patient, results in inefficient use of resources, and reduces throughput [42]. Thus the bed assignment process involves the simultaneous consideration of objectives to improve timeliness of assignments as well as appropriateness of assignments. The metrics by which bed management performance is benchmarked nationally are the number of diversions due to capacity restrictions and the median time between request and assignment. Low mean and median assignment times represent the reduction of system bottlenecks, improvements in timeliness of care, and efficient use of hospital resources. In addition to these benchmarks, we also seek to reduce the rate of off-service assignment and patients leaving without assignment.

1.2.2 Methodology

We describe the bed assignment problem using a queueing framework to account for the interactions between capacitated hospital units and the impact of boarding in creating throughput bottlenecks. Utilizing this queueing framework, we formulate a mathematical model of the system as a Markov decision process (MDP). Due to the complexity of the various streams of information and the detailed nature of the MDP state space, identifying optimal state-dependent routing policies can be difficult even for small hospital systems. Hence, we develop three assignment algorithms including those that account for probabilistic expectations about future events (e.g., arrivals and discharges) and utilize stochastic optimization methods. We test the performance

of these algorithms against approximations of current hospital practices and provide an upper bound on the performance using a simulation approach.

1.2.3 Contribution

Decision making related to bed assignments, which is a critical process impacting patient care and system efficiency, is a complex combinatorial optimization problem due to the dynamic nature of hospitals including the constant change in status and expectations for patient arrivals, transfers, and discharges. The problem we address is different from previous research on dynamic assignment problems [76] due to the dynamically changing information about both supply and demand for resources, the reuse of resources, and the expectations about future events. While other researchers have attempted to model the impact of bed assignment overflow policies [80] and blocking in hospital systems [11] to our knowledge no one has specifically examined the performance of bed assignment policies, incorporating the costs of boarding and value of the appropriateness of the assignment.

Through the development and testing (via simulation) of bed assignment algorithms we study the impact of how expectations about future events can improve bed assignment practices and hospital system performance. Additionally, we demonstrate the value of implementing strategic bed assignment techniques in a variety of hospital settings. These techniques are shown to balance the multiple objectives considered by bed managers, including both timeliness and appropriateness of bed assignments.

1.3 Patient Redirection in Hospital Networks

Building on the concept of strategic assignment to "overflow" beds, in Chapter 4 we examine the development of admission control policies for internal hospital units. These admission control policies allow for the redirection of patients to partner hospitals, if the unit to which they are seeking admittance does not have beds available.

We examine the development of patient-type specific redirection policies for a

particular unit. Policies take the form of defining thresholds for each of multiple patient types, such that if the number of patients in the unit reaches the threshold level the corresponding patients are redirected.

We assume a penalty is incurred while patients are redirected. These penalties or costs are assumed to be dependent on the patient type and represent the costs associated with the quality of care provided to a patient. This stems from the fact that often the appropriate equipment or services may not be available at another hospital, resulting in a higher cost for redirection.

We assume that the patient types (and their corresponding redirection costs) are ordered based on the hospital's priorities, such that one type of patient with a higher priority (and redirection cost) cannot be redirected unless patients with lower redirection costs are also redirected. Thus, we seek to identify threshold policies for patient redirection such that the prioritization and preferences for redirection of different patient types is conserved.

1.3.1 Objectives

With the consideration of multiple patient types we account for the appropriateness of the patient-unit through the redirection costs. We seek to minimize the sum of the average (per unit time) redirection costs incurred by all patient types.

We consider the hospital's objectives of efficiently utilizing resources while also minimizing the probability that a unit is completely full and cannot accept any incoming patients. We define the diversion rate to be the probability that the unit is completely full and therefore must divert, or redirect, all incoming patients. Additionally, we consider efficiency by calculating the long-term average utilization for the corresponding unit.

In choosing thresholds for a redirection policy we seek to (i) minimize the average cost from redirection, (ii) minimize the diversion rate, and (iii) maximize the average

utilization. Due to the nature of these performance metrics, tradeoffs in the objectives must be considered in the development of a patient redirection policy.

1.3.2 Methodology

With the assumption of preferences among the patient types, we model the unit as a loss system with the assumption of Poisson arrivals and exponential service times. For a particular policy, defined by the choice of redirection thresholds, the system is modeled as a Markov chain. Hence, for each policy the limiting probabilities for the number of occupied beds, the average cost of diversion, and the average utilization is calculated directly.

Utilizing the properties of the model, we demonstrate the impact of incremental changes to the thresholds of a control policy on the performance with respect to the three objectives defined above. As a result, we derive necessary conditions for an optimal threshold policy that minimizes the average cost per unit time. We conduct numeric experiments on sample instances to illustrate the validity of the necessary conditions for an optimal policy, the impact of preference constraints, and the tradeoffs in the three objectives.

1.3.3 Contribution

Unlike previous research, we examine the development of multiple threshold policies in which preferences among patient types are fixed and demonstrate the nature of tradeoffs in multiple objectives. Previous research either considers control systems with at most two thresholds, does not enforce preference constraints, or utilizes different objective functions. Our model can be applied to settings with any number of patient types and we demonstrate necessary conditions for a threshold policy that minimizes average cost of redirection with the assumption of preferences for patient types.

1.4 Outline

The remainder of this dissertation presents the methodology and policy development for multi-objective resource allocation problems in the three health care settings defined above. Chapter 2 describes the model of health care distribution in African countries, defines the hierarchical optimization approach used in policy development, and examines the tradeoffs in the three objectives and the impact of infrastructure characteristics.

Chapter 3 presents the queueing and MDP formulations of the bed assignment problem, the bed assignment algorithms which incorporate the definition of these models, and a study of the performance of the algorithms under realistic scenarios informed by data from a local hospital.

Chapter 4 defines the problem of identifying patient type-specific redirection policies for an internal hospital unit with redirection preference constraints, examines structural properties of the model, and presents an analysis of tradeoffs in multiple objectives with a computational study.

CHAPTER II

HEALTH CARE ALLOCATION IN SUB-SAHARAN AFRICA: TRADEOFFS IN EQUITY, EFFICIENCY, AND EFFECTIVENESS

In this chapter, we examine the allocation of scarce resources in African health care delivery. In particular, we utilize a hierarchical optimization approach to define resource allocation strategies in response to system constraints, distribution of population, and infrastructure characteristics. We study the impact of these features and characterize the tradeoffs in three system objectives: equity, efficiency, an health impact under different resource allocation strategies.

The remainder of the chapter is organized as follows. Background information regarding the problem setting is presented in Section 2.1. A description of the model, including assumptions, is provided in Section 2.2. Methodology for the creation of the instances used in the computations study and the hierarchical optimization approach, are explained in Section 2.3.1 and 2.3.2, respectively. An analysis of the results is presented in Section 2.4 including: a study of the impact of the prioritization of health and equity objectives (Section 2.4.1), a comparison of equity metrics (Section 2.4.2), an analysis of tradeoffs in health and equity (Section 2.4.3), and an examination of the impact of infrastructure (Section 2.4.4 and Section 2.4.5).

2.1 Introduction and Literature Review

In many parts of the developing world, including much of the African continent, there is a crisis because people lack access to essential medicines and health services. The severe imbalance of supply and demand for medical care dramatically increases the

complexity of the allocation decisions that have to be made by many ministries of health (MOHs) and non-governmental organizations (NGOs).

Additionally, much of the difficulty in allocating scare resources arises from the many competing objectives that must be considered, e.g., maximizing health impact, minimizing resource usage, and ensuring fair allocations. These objectives of effectiveness, efficiency, and equity are key elements in the delivery of quality health care in Africa [8, 85]. Health care delivery in resource-constrained regions is further complicated by unreliable transportation systems, the relative locations of health centers and settlements, and diverse geography [78, 85]. Therefore, health care allocation decisions must account for the logistical challenges resulting from the location of populations and health centers as well as the configuration and quality of the transportation infrastructure.

These logistical challenges directly impact the effectiveness, efficiency, and equity of an allocation strategy. For example, due to the spatial characteristics of a region, it may be least expensive to the health care system to distribute health services in densely-populated areas with communities in sparsely populated areas incurring an unreasonable burden to receive care. Increasing equity by delivering health care to those in remote areas will require a sacrifice in efficiency, with a significant increase in the average cost per patient treated [24]. Due to this increase in average patient cost, limited funds would treat fewer patients thereby decreasing the overall effectiveness of the allocation.

No consensus has been reached regarding the appropriate definition and measurement of equity [65]. We use two objectives to achieve equity between population centers or communities. The first is minimizing the difference in the health improvement in the communities with the largest and smallest improvement. The second is maximizing the health improvement at the population center with the smallest improvement. This second objective incorporates Rawlsian's maximin concept, by which

resources are used to improve the status for the worst off-individual, or community [65]. Both objectives strive for vertical equity, in which resources are distributed to locations according to the relative needs [65].

To examine tradeoffs in effectiveness, efficiency, and equity we develop a mixed-integer program (MIP) to model an African health care system and the corresponding decisions for the allocation of scarce resources. Mixed-integer programming is a technique that has been utilized in other health care allocation models for resource-constrained countries, including models which efficiently allocate resources among different treatments while comparing the performance of multiple objectives [23]. Unlike these health care allocation models, we incorporate multiple objectives in a hierarchical fashion.

Most models for health care allocation in Africa focus only on antiretroviral HIV treatments [44, 86]. Some HIV treatment allocation models incorporate spatial characteristics of specific African regions [14, 83, 84]. Wilson and Blower (2005) developed a mathematical model for the equitable allocation of antiretrovirals in South Africa, which incorporated the relative locations of patients and health centers. The authors identified an allocation strategy which maximized equity in the percentage of infected individuals receiving treatment in each community [83]. Unlike Wilson and Blower, our model incorporates transportation and distribution costs, and dependencies on the infrastructure system.

Rather than focusing on one specific region in Africa, as many other studies have done, we examine a variety of settlement patterns often seen in Africa. We implement different allocation strategies for all settlement patterns, where an allocation strategy represents choices regarding the importance assigned to health outcome and equity.

Our computational study reveals that a careful choice of an allocation strategy is important in environments with tightly constrained budgets. The settlement pattern of the environment seems to have a smaller impact on which allocation strategy is

best. When limited funds are available, significant differences in health outcome and equity can occur for the different allocation strategies. We find that for each settlement pattern and each budget level, there is a specific "equity-threshold" at which a minor increase in the total health improvement requires a large sacrifice in equity.

We also conducted a comprehensive study to understand the impact of the location of health centers and the quality of the transportation infrastructure on the efficiency of the delivery of health care. We find that the efficiency does depend strongly on the settlement pattern. Furthermore, we see that targeted investments in infrastructure improvement are far more effective than system-wide infrastructure improvements (for a given investment budget).

2.2 Model

Health care allocation systems in Africa are extremely complex involving interactions between patients, health workers, transportation systems, health centers, and political agencies. The model we develop incorporates many of these complexities while also making some simplifying assumptions.

The model takes as input information pertaining to the settlement pattern and the health care and transportation infrastructure of a particular region, including a set of population centers (PCs), a road network connecting these population centers, a set of health centers (HCs), and a medical supply distribution network connecting health centers. Each population center is characterized by its total population, or census, and disease occurrence. Rather than limiting the study to only one disease, we allow multiple diseases each with its own occurrence rate. A road is defined by the two population centers it connects, its length, and its quality. To accommodate prevailing conditions within Africa, roads are categorized as paved roads, dirt roads, or foot paths. The quality of a road is classified as primary, secondary, or tertiary,

respectively. The transportation costs and travel time of a road depends on type and length.

Reflecting health care systems in Africa, each health center is classified as a regional hospital or a rural health center, defining the level of care it provides. These are referred to as primary-level and secondary-level health centers, correspondingly. In addition to defining the level of service, this classification also identifies the role of the health centers in the distribution of medical supplies. Often, in practice, medical supplies are distributed in a top-down fashion from intermediate stores located at regional hospitals to rural health centers [1, 69]. This hierarchical distribution system is incorporated into the model, with the cost of distributing from one health center to another dependent on the roads that are traveled in the shortest path between the corresponding population centers.

The model determines the allocation of medical treatments to health centers as well as the allocation of medical treatments to population centers. The model assumes that each disease has a particular treatment and that all treatments are administered in the communities by visiting nurses, or health workers, dispatched from the nearest open health center. A health center is open if it is allocated any treatments. Without explicitly allocating treatments to population centers (through the assignment of health workers to population centers), the percentage of treated individuals at a population center would be more difficult to establish. It would depend on modeling the willingness of individuals to travel to the nearest open health center and on modeling how the limited number of medical treatments available at open health centers would be distributed among the individuals seeking treatment. Thus, we model delivery of care, rather than access to care. This allows for a more controlled study of equity in health care provision. In addition, it also reflects current practice as many organizations have been working on improving their ability to send health workers to patients, including Riders for Health [79].

The number of individuals that can be treated in one day by a nurse from a particular health center depends on the time required to administer treatments and the nurse's time spent traveling between the health and population centers. These travel times are impacted by the transportation network and road conditions. Each nurse can only visit one population center in a day and we define a capacitated quantity of nurse-days at each health center. Each nurse-day represents one health worker being available for one day. Since treatments can only be administered by nurses, the number of treatments allocated to a population center, is affected by the number of nurses at the nearest open health center. This limit on the staff capacity represents of one of the major challenges in health care delivery in Africa: the significant shortage of health workers [15, 85]. All patients with an illness from a particular population center are assumed to be available for treatment during a nurse's visit.

The effectiveness of an allocation policy is measured by the health improvement achieved. Health improvement is typically measured in quality-adjusted life-years (QALYs) [7]. In our model, each treatment is assigned a health improvement value, measured in QALYs, representing the expected impact of the treatment. We assume that health improvements (QALYs) are additive if a patient is treated for multiple illnesses. Each treatment is also assigned a purchase cost. Therefore, the health care budget is spent on purchasing medical treatments as well as on shipping treatments to health centers through the distribution network.

We employ four objectives to define allocation strategies. These objectives pertain to effectiveness, efficiency, and equity. The first objective relates to effectiveness and maximizes the total health improvement, or the total increase in QALYs achieved in the total population. Efficiency is achieved by fixing the level of health improvement and minimizing the total cost, where total cost includes the cost of opening of health

centers, the cost of distribution, and the cost of procuring medical treatments. Therefore efficiency is measured as the cost per QALY ($/QALY). The final two objectives relate to equity. They are both expressed in terms of the *success rate* at population centers, or communities. The success rate at a population center is the ratio of the health improvement achieved by an allocation strategy and the maximum possible health improvement, i.e., the health improvement when every individual in the population center is treated. The first equity objective maximizes the minimum success rate over all population centers. The second equity objective minimizes the difference between the smallest and largest success rates among the communities.

We developed a mixed-integer program (MIP) to mathematically model and evaluate allocation strategies. The MIP captures the relationships and interactions between transportation infrastructure, health centers, health workers, communities, patients, and health treatments. There are constraints that limit the number of patients treated based on the supply and demand for nurses and treatments. Supply levels are determined by the number of treatments allocated, the number of nurse-days at open health centers, and nurse traveling distances. Patient demand at an open health center depends on which other health centers are open since patients receive treatment from the nearest open health center. There are constraints that capture the structure of the distribution network and ensure that treatments are shipped from regional health centers to rural health centers. Finally, there is a budget constraint that limits the expenditures. The full formulation and description of the model is included in Appendix A.

This model is most applicable to the allocation of treatments for diseases that require regular treatment and in which the disease occurrence can be estimated. Examples include HIV treatments, distribution of contraceptives, deworming medicine, vaccinations, palliative care, or the treatment through bed nets. In such cases, decisions for the allocation of treatments can be repeated at regular intervals.

2.3 Methodology

2.3.1 Data Creation

One of the significant challenges associated with research related to health care allocation in Africa, especially when taking infrastructure characteristics into account, is the absence of reliable data. Lacking reliable data sources, we have developed a methodology to create data instances that mimic health systems in Africa. The ability to generate and analyze many data instances allows for a more robust study of the behavior of allocation strategies in areas with a variety of spatial characteristics.

To create a plausible representation of common settlement patterns, a set of population centers, or communities, was created, each classified as either rural or urban. The census at a population center was drawn from a distribution dependent on its classification, with a significantly higher average census in urban centers than in rural ones.

Population density and settlement patterns throughout Africa can vary greatly depending on the country or location on the continent [16]. Stock (2004) noted that a variety of rural settlement patterns exist in Africa including clustered, or nucleated, settlements in the form of villages [78]. When not clustered, rural settlements in Africa are generally dispersed throughout a region. To account for some of the variability in spatial patterns, we created instances of three different types. These types reflect both rural settlement patterns and the relative location to the urban population centers. The three spatial types are characterized as "rural population centers sparsely distributed away from urban centers," "rural populations clustered away from urban centers," and "rural population centers clustered near urban centers".

Since roads generally connect cities that are near to each other rather than cities that are far apart, it is assumed that roads only exist between population centers that are less than a predetermined distance apart. The probability of a road existing and its type depend on the classification of the two corresponding population centers

as urban or rural and the distance between the population centers.

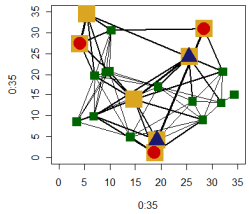
Additionally, each instance definition includes the location of health centers at urban population centers. A health center is designated as either primary-level or secondary-level according to a probability distribution. At most one health center is located at an urban population center and health centers have a higher probability of being located in those urban population centers in which the population density of the surrounding area is greater. An equal number of nurse-days is allocated to each health center.

Table 1: Input parameters

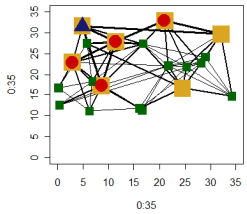
Data Generation Input	Value
Total Area (sq. miles)	4900
Number Population Centers	20
Average Population - Urban PC	1000
Average Population - Rural PC	200
Incidence Rate - Disease 1	20%
Incidence Rate - Disease 2	10%
QALY Improvement per Treatment - Disease 1	.225
QALY Improvement per Treatment - Disease 2	.525
Treatment Cost - Disease 1 ($)	20
Treatment Cost - Disease 2 ($)	35
Number of Health Centers	5
Probability of Primary Health Center	.4
Cost of Opening a Health Center ($)	1000
Number of Nurse-Days Per Health Center	10
Time Per Nurse-Day (hours)	6
Nurse's Maximum Travel Time (minutes)	80
Time Per Treatment (minutes)	5
Nurse Travel Speed on Level 1 Road (mph)	40
Nurse Travel Speed on Level 2 Road (mph)	27
Nurse Travel Speed on Level 3 Road (mph)	20
Transport Cost Per Mile Per Treatment on Primary Road ($)	1.25
Transport Cost Per Mile Per Treatment on Secondary Road ($)	1.75
Transport Cost Per Mile Per Treatment on Tertiary Road ($)	2.25
Probability of Road Between Two Urban Centers Being Primary	0.5
Probability of Road Between Two Urban Centers Being Secondary	0.5
Probability of Road Between One Rural and One Urban PC Being Secondary	0.5
Probability of Road Between One Rural and One Urban PC Being Tertiary	0.5
Probability of Road Between Two Rural Centers Being Tertiary	1

Using a consistent set of parameters, shown in Table 1, 90 data instances were created for the initial analysis using this methodology. Each of the three spatial

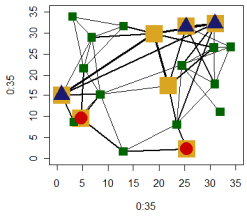
patterns accounted for one-third of the data instances. Thirty infeasible instances, in which not all nodes were connected, were excluded from the analysis. A visual representation of some of the data instances is shown in Figure 1. In the following sections, we highlight some of these instances to demonstrate key behaviors that are exhibited throughout the set of data instances.



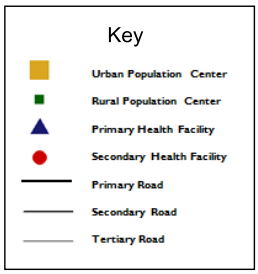
(a) Rural PCs Clustered Away From Urban PCs



(b) Rural PCs Clustered Near Urban PCs



(c) Rural PCs Scattered Away From Urban PCs



(d)

Figure 1: Instance examples for three spatial patterns

2.3.2 Multi-Objective Optimization

Allocation strategies are constructed using efficiency, effectiveness, and equity objectives. Each allocation strategy is defined by an objective hierarchy, which ranks the objectives in order of importance. Hierarchical optimization is then used to find the actual allocation for a particular data instance. More specifically, we start by optimizing with respect to the objective at the top of the hierarchy. When the optimal value for that objective has been found, a constraint is introduced in the optimization

model to ensure that the value of that objective is no worse in any subsequent optimizations. Next, we optimize with respect to the second objective in the hierarchy. As before, when the optimal value for that objective has been found, a constraint is introduced in the optimization model to ensure that the value of that objective is no worse than in the last optimization. Finally, we optimize with respect to the third and last objective in the hierarchy.

Four allocation strategies are considered using this hierarchical optimization approach differing in the first two objectives in the hierarchy: WSHI, DSHI, HIWS, and HIDS. The last objective is always minimizing cost. Table 2 defines the ordering of the objectives for each of these four strategies.

Table 2: Definition of allocation strategies

Allocation Strategy	First Objective	Second Objective
WSHI	Max Worst Success Rate (Equity)	Max Health Improvement (Effectiveness)
DSHI	Min Difference in Success Rates (Equity)	Max Health Improvement (Effectiveness)
HIWS	Max Health Improvement (Effectiveness)	Max Worst Success Rate (Equity)
HIDS	Max Health Improvement (Effectiveness)	Min Difference in Success Rates (Equity)

Alternatively, the multiple objectives could have been combined into a single objective using a weighted sum. The disadvantage of using a weighted sum is the need to define the weights. For our model, defining weights would be difficult due to the inconsistencies in the units of measure, including health improvement (QALYs), success rate (%), and cost ($). With a hierarchical optimization approach the tradeoffs in effectiveness, efficiency, and equity can be studied without the definition of weights.

2.4 Analysis

To analyze the performance of the allocation strategies (Section 2.4.1), the choice of equity objective (Section 2.4.2), and to better understand the characteristics of the fundamental tradeoffs (Section 2.4.3), each allocation strategy was applied to each of the 60 data instances. As shown in Table 3, the potential health improvement is similar for each of the spatial distribution types.

Table 3: Health improvement (in QALYs) among all feasible instances

Spatial Distribution Type	# of Instances	Minimum	Maximum	Mean	Median
Rural PCs Clustered Away From Urban PCs	18	746	2044	1069	903
Rural PCs Clustered Near Urban PCs	22	793	1816	1136	917
Rural PCs Scattered Away From Urban PCs	20	761	1950	1076	882
All	60	746	2044	1096	901

However, as shown in Table 4, the cost of health improvement, in dollars per QALY, assuming a budget of $100,000 is quite different for the allocation strategies. On average the cost per QALY improvement is much greater for strategies that prioritize equity over health, i.e., WSHI and DSHI, than for the strategies that prioritize health over equity, i.e., HIWS and HIDS. While this basic analysis provides insight into the relative behavior of the strategies, a closer examination of the behavior and other metrics provides a more complete understanding of the impact of budget, choice of equity objective, and fundamental tradeoffs.

Table 4: Cost per QALY ($/QALY) for allocation strategies WSHI, DSHI, HIWS, and HIDS: $100,000 Budget

Allocation Strategy	Minimum	Maximum	Mean	Median
WSHI	17	222	137	131
DSHI	17	222	143	141
HIWS	67	172	109	107
HIDS	67	172	110	108

2.4.1 Impact of Budget When Prioritizing Health and Equity

In sub-Saharan Africa, limited budgets for health care are common. Therefore, understanding the impact of budget on allocation strategies is important. For each instance and strategy, we identify the optimal allocation at many budget levels. The value of an allocation is measured in total health improvement (QALYs) over all population centers, the difference in success rates at the best- and worst-off population centers, and the minimum success rate among population centers. A more informative comparison of the value of allocation strategies is obtained by examining allocations for different budget levels.

For Instance 9, a graphical representation of the performance of the allocation strategies is shown in Figure 2. Each chart represents the allocations obtained with a particular strategy for different budget levels. The minimum success rate among all population centers for an allocation is represented by the striped bar. The difference between the minimum and maximum success rates among all population centers for an allocation is represented by the height of the second bar. Consequently, the maximum success rate among all population centers for an allocation is displayed as the top of the stacked bars.

We see that for Instance 9, the ranking of objectives can significantly impact the value of an allocation. This is demonstrated most clearly in the comparison of allocations obtained with a budget of \$75,000. When health improvement is the primary objective and equity is sought by maximizing the worst success rate (Fig. 2(c)), 838 QALYs are achieved with some population centers receiving no treatment and some population centers receiving full coverage. However, when the order of the objectives is reversed, i.e., when equity is prioritized over health outcome(Fig. 2(a)), only 622 QALYs are achieved, with all population centers achieving at least a 71.6% success rate and the best off population center achieving a success rate of 96%. This demonstrates that the quality of an allocation can be significantly different for different allocation strategies, i.e., for a different ranking of objectives.

The greatest difference between strategies WSHI and HIWS, or between DSHI and HIDS, occurs at low budget levels. At these levels, the difference between the best and worst success rates is larger and the worst success rate is smaller when health improvement is the primary goal. For example, in Instance 9 with a budget of \$75,000, when equity is the primary objective and sought by minimizing the difference between best and worst success rates(Fig. 2(b)), each population center achieves a success rate of 71.6%. By contrast, some population centers do not receive any treatment when health is the primary objective(Fig. 2(d)). With limited budgets, emphasizing

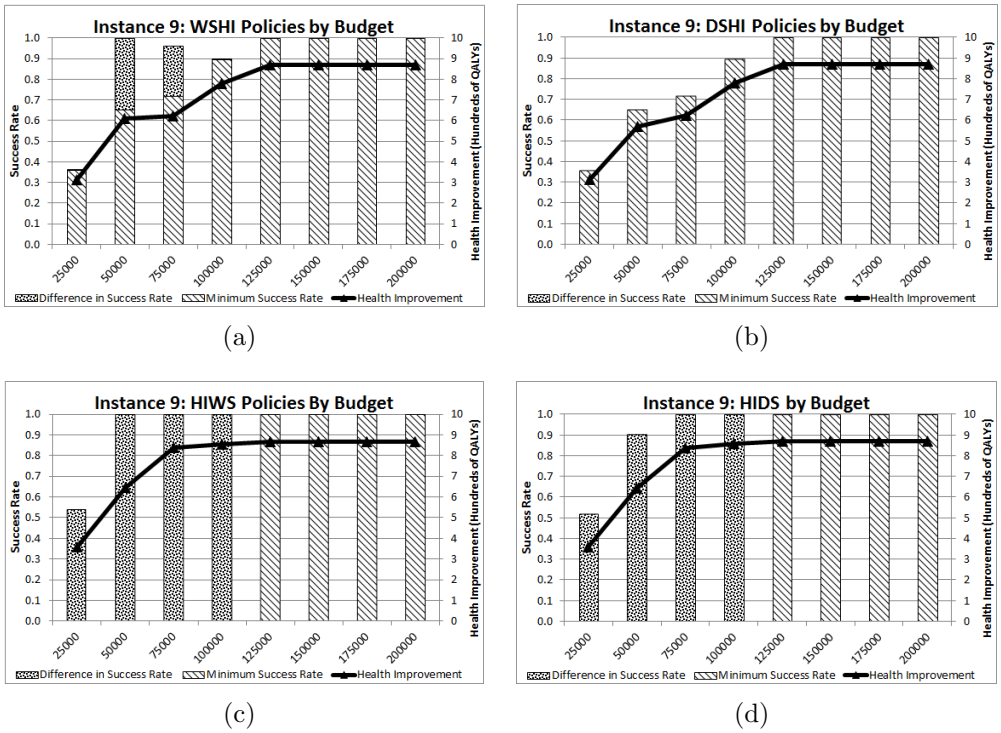


Figure 2: Performance of (a) WSHI, (b) DSHI, (c) HIWS, and (d) HIDS policies by budget [Instance 9]

health over equity results in allocations that are significantly less equitable than the allocations in which equity is preferred over health, regardless of how we seek to achieve equity. It is at these lower budget levels, i.e., when resources are scarce, that difficult allocation decisions are made in practice. Therefore the choice of the primary objective significantly impacts allocations in real world settings. Tradeoffs between health and equity are examined more closely in Section 2.4.3.

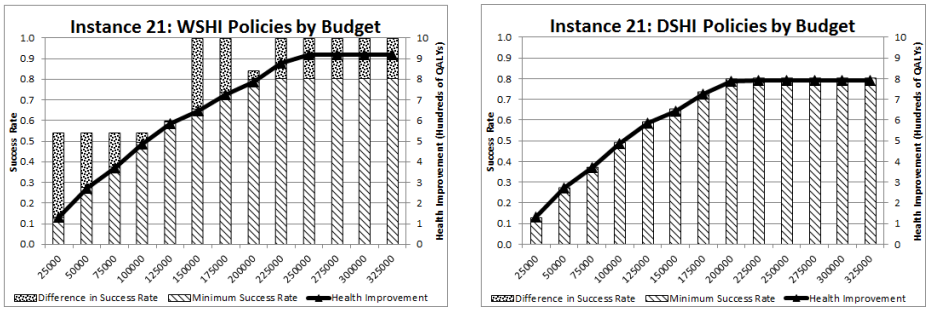
2.4.2 Comparison of Equity Objectives

To identify the impact of the equity objectives on the allocations, we contrast the WSHI and DSHI strategies, in which equity is prioritized, and the HIWS and HIDS strategies, in which health improvement is the primary objective.

In an allocation produced by the DSHI strategy all population centers will have the exact same success rate and the difference in success rates will be zero. Subsequently maximizing the health improvement with the difference in success rate constrained

to be less than or equal to zero (i.e., constrained to be equal to zero) is equivalent to maximizing the smallest success rate. Therefore, an allocation produced by the DSHI strategy will be optimal with respect to both equity objectives, but the health improvement will be less than or equal to that of the allocation produced by the WSHI strategy (assuming the same budget).

Large differences in health improvement in allocations produced by the WSHI and DSHI strategies occur when the budget is large and an increase of budget will not cause an increase in the minimum success rate. As seen in Figure 3, in Instance 21 the health improvement is similar for the allocations produced by the DSHI and WSHI strategies for budgets less than or equal to $225,000, but differences occur when the budget is greater than $225,000. For some instances and certain budgets, the WSHI and DSHI strategies result in allocations with a similar health improvement, but where the allocation produced by the WSHI strategy has a greater success rate difference, implying a less equitable allocation. In Instance 21, with budgets between $25,000 and $200,000 the success rate difference in allocations produced by the WSHI strategy is significantly greater than of those produced by the DSHI strategy, but the difference in health improvement is less than 1%. Therefore, seeking equity by minimizing the difference in success rate results in better allocations, with respect to health improvement and equity.

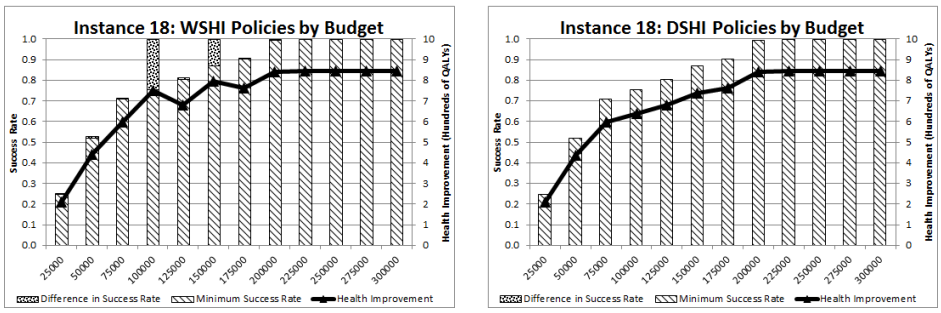


(a) (b)

Figure 3: Performance of (a) WSHI and (b) DSHI policies by budget [Instance 21]

There are instances and budget levels for which the allocation produced by the

WSHI strategy has greater health improvement and greater difference in success rate than the allocation produced by the DSHI strategy. In Instance 18, as shown in Figure 4, this occurs at budgets \$100,000 and \$150,000. With the \$100,000 budget, the allocation produced by the DSHI strategy has 15% fewer QALYs than the allocation produced by the WSHI strategy. For the other budgets, there is little difference in health improvement of the allocations produced by the strategies. Overall, the health improvement is comparable for allocations produced by the WSHI and DSHI strategies, but the DSHI strategy results in the more equitable allocations.

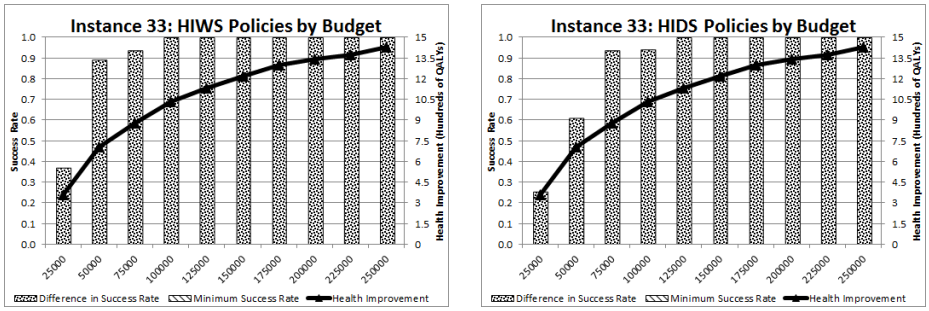


(a) (b)

Figure 4: Performance of (a) WSHI and (b) DSHI policies by budget [Instance 18]

We also contrasted the allocations resulting from strategies HIWS and HIDS to examine the effect of equity objectives when prioritizing health improvement over equity. For most instances, the allocations from strategies HIWS and HIDS are very similar. This is not surprising, because when health improvement is constrained to be maximal, there are typically only a few feasible allocations. Therefore, there is often little difference in allocations produced by strategies that optimize health first. With some instances and budgets there are substantial differences between the allocations produced by the two strategies. Usually this occurs when, due to the characteristics of the instance, it is impossible to allocate any treatments to a particular population center while simultaneously achieving optimal health improvement. In Instance 33, shown in Figure 5, the minimum success rate is zero regardless of budget size, but the success rate difference (or maximum success rate in this situation) is smaller in

the allocation produced by the HIDS strategy than in the allocation produced by the HIWS strategy.



(a) (b)

Figure 5: Performance of (a) HIWS and (b) HIDS policies by budget [Instance 33]

In most instances in which at least one population center does not receive any treatments, this is due to the limited number of health workers at the nearest health center. Under the assumption that all population centers are served by the closest open health center, opening a new health center will not always relieve this workforce shortage. While this assumption may seem limiting, it is not uncommon in developing countries of the world that shortages in health workers and the location of health centers result in some population centers not receiving health services. Therefore, these particular instances depict the challenges faced in practice when staffing resources capacity can be more limiting than financial resources.

Overall, we find that the HIDS strategy tends to produce more equitable allocations than the allocations produced by the HIWS strategy. Thus, we conclude that for the majority of instances and budgets, regardless of the ordering of the objectives, seeking equity by minimizing the difference in success rates should be preferred over seeking equity by maximizing the worst success rate.

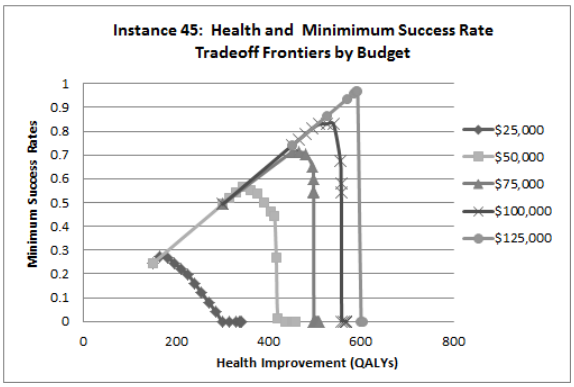
2.4.3 Tradeoffs in Health, Equity, and Budget

The analysis to this point has focused only on the allocations that can be obtained with the hierarchical optimization. We have seen that the ordering of the objectives

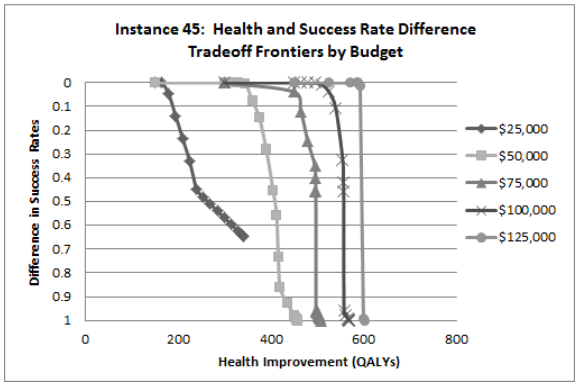
can significantly impact the value of the allocations obtained, often resulting in a significant tradeoff in equity and effectiveness. A deeper understanding of these tradeoffs is obtained by examining the tradeoff curve for a given budget. The points at the extremes of a tradeoff curve correspond to the allocations produced by the strategies introduced earlier. The other points on a tradeoff curve correspond to allocations obtained when the health outcome is fixed at a level other than those seen at the extremes.

First, we focus on tradeoff curves that result when we seek equity by maximizing the worst success rate. An example is shown in Figure 6(a). For a given budget, the part of the tradeoff curve to the left of the maximum corresponds to solutions in which the success rate is limited by the health improvement level, not by the budget; in fact the budget is not fully utilized in these solutions. Because the budget is not used fully, it is possible to achieve a greater level of health improvement without sacrificing equity (i.e., it is possible to move to the right in the graph). Therefore, the solutions to the left of the maximum will never be implemented in practice.

We define the maximum on the tradeoff curve as the "equity threshold". Beyond this threshold, an increase in health improvement requires a decrease in equity, or worst success rate. In Figure 6(a), with a budget of \$50,000, the equity threshold occurs at a health improvement of 330 QALYs and a worst success rate of 56.7%. The steep slope of the tradeoff curve to the right of the equity threshold demonstrates that in order to achieve greater health improvement, the population center with the worst success rate must receive substantially fewer health treatments, resulting in a less equitable allocation. For each budget, the right-most point on the tradeoff curve corresponds to the policy in which health is prioritized over equity. In Figure 6(a), this corresponds to a worst success rate of 0% for all budgets. It is interesting to observe that for budgets greater than \$75,000, the health improvement at the equity



(a)



(b)

Figure 6: Tradeoff frontiers: (a) Health-minimum success rate tradeoff; (b) Health-difference in success rate tradeoff [Instance 45]

threshold is only slightly less than the maximum achievable health improvement, but is significantly more equitable, i.e., a small sacrifice in health outcome can result in significant gains in equity.

A similar analysis is performed when we seek equity by minimizing the difference between the best and the worst success rates observed at population centers (see Figure 6(b)). The tradeoff curves in Figure 6(b) show that it is possible to have an equitable solution, but that to achieve the maximum possible health outcome equity has to be sacrificed; in fact, in allocations that result in the maximum health outcome there is at least one population center that receives no treatment and at least one population center that receives treatments for all its patients, i.e., difference in success rate of 100%. Again, we observe that beyond a certain health outcome an increase

in health improvement comes at a high price in terms of equity.

The range of health improvement levels where a small increase in health requires a significant decrease in equity, referred to as the "critical health region", is virtually independent of the way we seek equity. As shown in Figure 7, there appears to be an an intrinsic range of health improvement levels at which the tradeoff in health improvement and equity is great, which depends on the instance and the budget. Furthermore, with larger budgets, the critical health region becomes narrower.

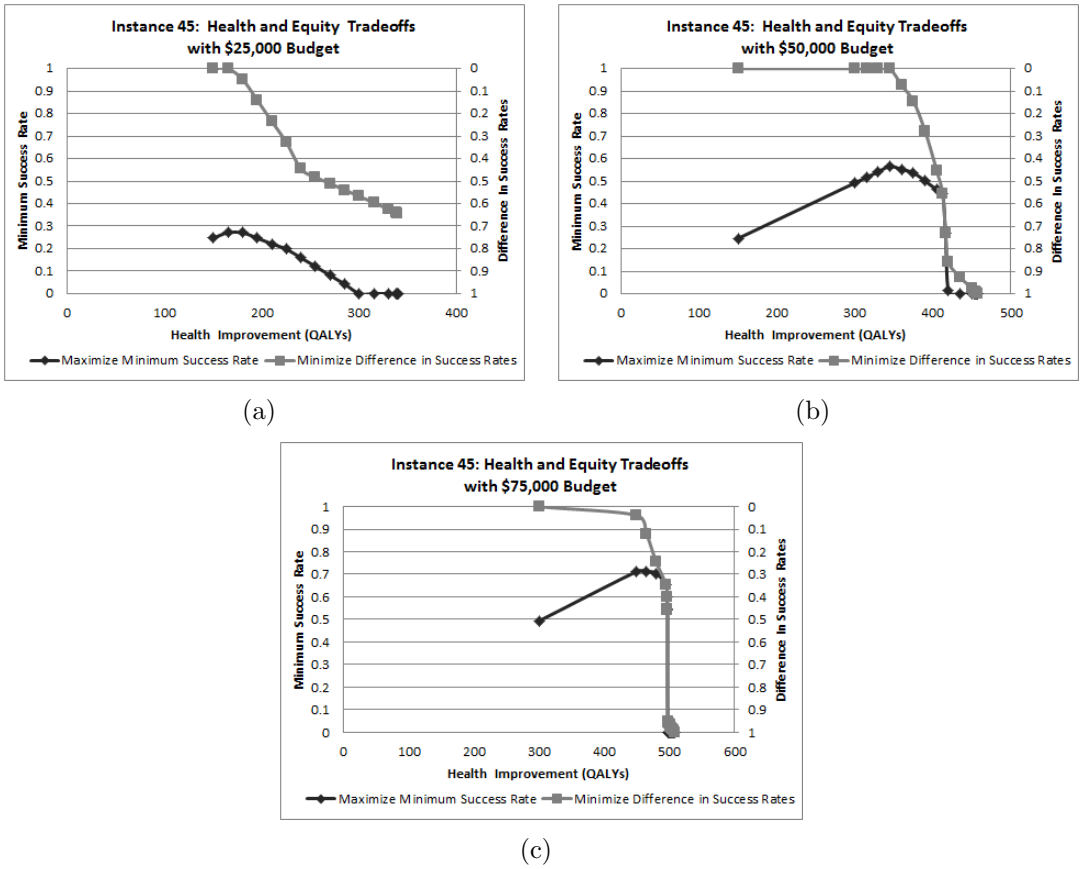


Figure 7: Health-Equity tradeoff frontiers with budgets of (a) $25,000 (a), (b) $50,000 (b), and (c) $75,000 [Instance 45]

We hypothesize that the critical health region depends on the unique characteristics of an instance, including health center and population center locations as well as characteristics of the road network, but not on the way equity is sought. The critical health region appears to be quite narrow, other than for very small budgets.

Therefore, determining and analyzing the critical health region is enormously valuable when making difficult allocation decisions.

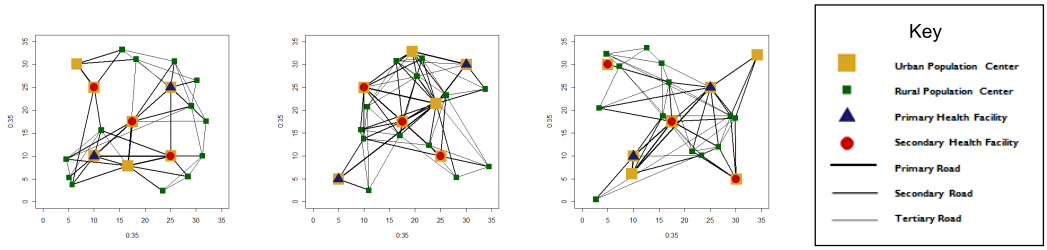
2.4.4 Impact of Road Infrastructure, Settlement Patterns, and Health Center Location on Efficiency

It is reasonable to assume that the efficiency of an optimal allocation is affected by the settlement pattern, health center locations, and road infrastructure. To study these influences, we created 60 additional instances and examined the impact of road infrastructure on the average cost per QALY when employing the DSHI strategy, where we measure the quality of road infrastructure as the percentage of the total road distance classified as primary road. We plot the primary road percentage against average cost per unit health improvement of an optimal allocation for a variety of instances and budgets. Each allocation is classified according to the overall success rate achieved. Therefore, each point in a plot represents the primary road percentage and cost per QALY for all allocations that achieve the corresponding success rate level.

To create a fair comparison, each instance consists of the same set of population centers, although the locations and road networks vary. To understand the impact of the location of health centers, the set of population centers where health centers are located is fixed as one of ten different patterns. By holding these characteristics of the population centers and the location of health centers constant, the differences between instances are related to the spatial distribution of population centers and the road infrastructure. Instances are created according to the three spatial distributions defined earlier: rural population centers clustered near urban population centers, rural population centers clustered away from urban population centers, and rural population centers sparsely distributed away from urban population centers.

First, we examine the impact of the location of the health centers on the relationship between the road infrastructure quality and the efficiency. Plots of the

relationship between road infrastructure quality and efficiency for the three health center location patterns are given in Figure 9. For the first pattern (Fig. 8(a)), in which all health centers are centrally located, the average cost per unit of health improvement does not appear to be impacted by the quality of the roads. However, for the second pattern (Fig. 8(b)), in which the distance of the primary health centers to the populations centers is larger, we see that as the quality of the roads increases the average cost per QALY decreases. This can be explained as follows. As the quality of the roads increases, nurses experience shorter travel times, allowing them to treat more patients. At the same time, distribution costs between health centers decrease, making the total cost less expensive. Both result in a decrease in the average cost per QALY.



(a) Centrally Located Health Centers

(b) Increased Distance to Primary Health Centers

(c) Increased Difference to Secondary Health Centers

(d)

Figure 8: Health center location patterns

A similar effect is observed with the third pattern (Fig. 9(c)), in which the distance of the secondary health centers to the population centers increases. Here we see that when the percentage of primary roads is below 4%, the cost per QALY ranges between \$100 and \$250 for instances in which success rates are between 65 and 75 percent. However, when the percentage of primary roads is greater than 4%, the cost per QALY is less than \$133 regardless of the success rate. This demonstrates that investment in road infrastructure can have a significant impact on the efficiency of a health allocation, but that the impact depends on the location of the health centers.

Next, we consider balanced and unbalanced locations of health centers. In a

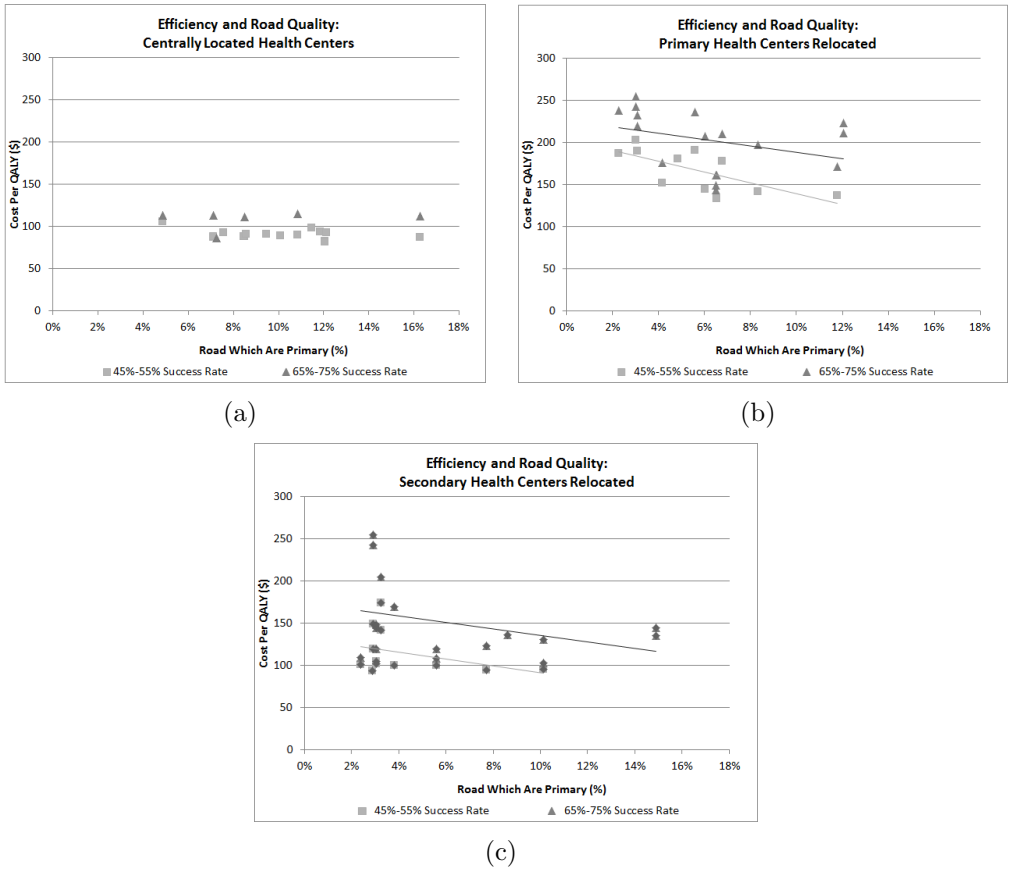
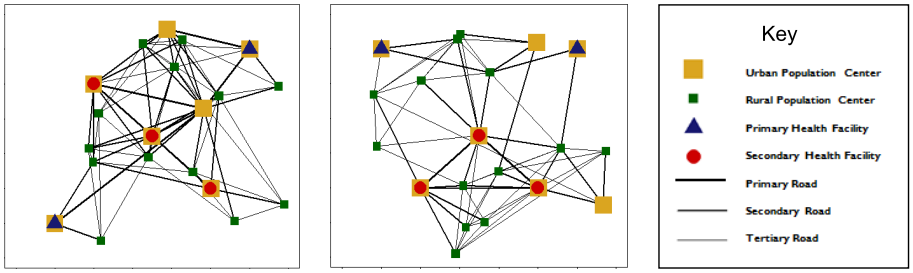


Figure 9: Efficiency and road quality for (a) centrally located health centers , (b) relocated primary health centers , and (c) relocated secondary health centers

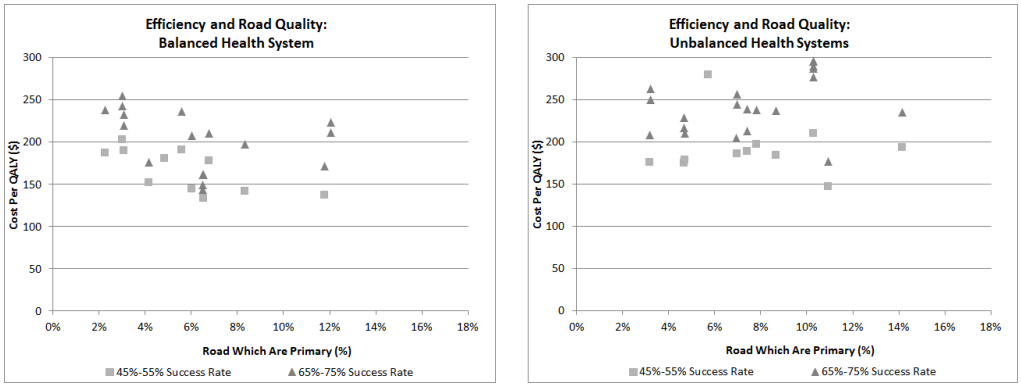
balanced health system, the primary health centers are located so that they have good coverage of the entire region (Fig. 10(a)). In an unbalanced health system, the primary health centers are located on one side of the region and do not have good coverage of the entire region (Fig. 10(b)). The plots in Figure 11 show similar trends as before, i.e., a reduction in cost per QALY when the road infrastructure improves, although not as pronounced. More importantly, the plots show that the cost per QALY to achieve a comparable health outcome is higher for an unbalanced system. This again underscores the significance of health center location in the efficiency of health distribution systems.

Lastly, we investigate the impact of the settlement patterns on the interaction of road infrastructure and cost efficiency. In Figure 12, we again plot the cost per QALY



(a) Balanced Location of Health Centers (b) Unbalanced Location of Health Centers (c)

Figure 10: Balanced and unbalanced health systems: health center location patterns



(a) (b)

Figure 11: Efficiency and road quality for balanced health systems (a) and Unbalanced health systems(b)

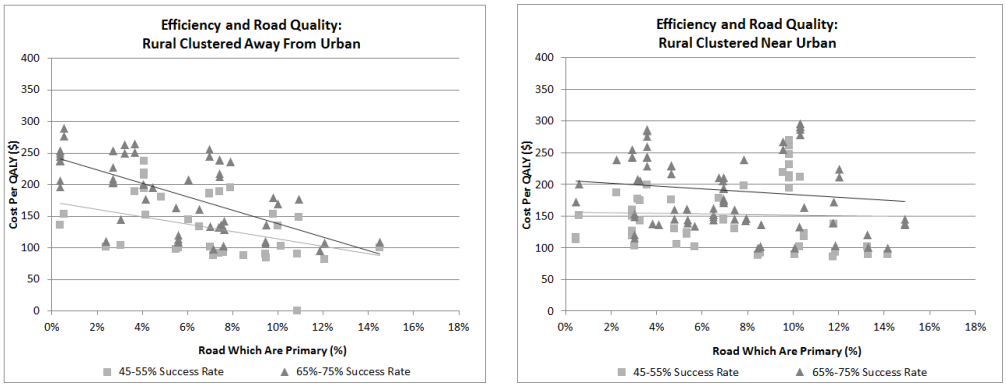
against primary road percentage for a given success rate level, but now the instances are divided according to their settlement pattern.

When rural population centers are clustered away from urban population centers, the cost per QALY shows a clear downward trend when road quality increases. The trend is less clearly discernible for the other two settlement patterns.

2.4.5 Impact of Infrastructure Improvements

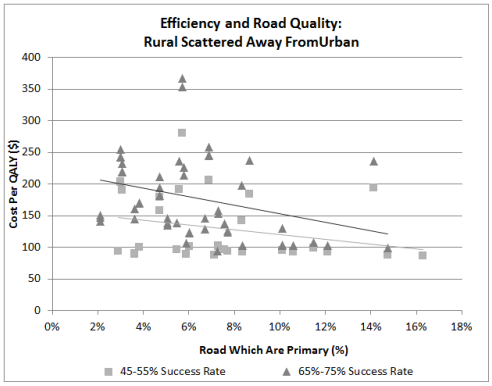
To further evaluate the impact of road improvements on the efficiency of health care delivery, we created two sets of eight instances where the only difference between the instances in a particular set is the quality of the roads.

For the first set, we observe, as expected, that the cost per QALY decreases when



(a)

(b)



(c)

Figure 12: Efficiency and road quality for spatial patterns: rural clustered away from urban (a), rural clustered near urban (b), and rural away (c)

the quality of the roads increases. A more detailed analysis reveals additional insights. The three instances with the lowest cost per QALY not only had the highest primary road percentages, but also all had a primary road connecting an isolated primary health center with a centrally located secondary health center. In instances with similar primary road percentages, but in which this particular road was not a primary road, the cost per QALY was higher.

For the second set, we did not observe that the cost per QALY decreases when the quality of the roads increases. In fact, there are two instances in which only 10% of roads are primary but that have a significantly lower cost per QALY than instances in which 18-19% of the roads are primary. Upon closer examination of the instances in this set, it was found that three of the roads played a key role in the

efficiency of allocations. Instances in which at least one of these three roads was primary had a low cost per QALY where instances in which none of the three roads were primary had a high cost per QALY. Two of these roads connected primary and secondary health centers. The third road was between a secondary health center and an isolated rural population center. However, the census at this isolated location is largest among the rural centers, and this road is the only connection to the nearest health center. Therefore, despite the quality of the road it is used frequently. As a result, an improvement to this road has an immediate and noticeable impact on the efficiency.

From the analysis of both sets of instances, it is apparent that targeted road improvements are necessary to increase the efficiency of allocations. This also implies that an investment in infrastructure improvement does not have to be large to have an impact on the efficiency of health care delivery.

2.5 Conclusion

By analyzing and comparing health care allocations in a large number of situations with characteristics that are representative of environments encountered in regions of sub-Saharan Africa, we were able to understand and evaluate the tradeoff in equity and effectiveness as well as the impact of health care and transportation infrastructures on health outcome.

The major findings of our study can be summarized as follows:

- Decision makers should aim to distribute medical supplies and treatments in proportion to the relative needs of population centers if they want to ensure an equitable delivery of health care.

- Tradeoff curves of health outcome and equity indicate a clear structure that shows that there is a point after which even a small increase in health outcome

leads to a large decrease in equity. As a consequence, there is only a small set of health care allocations that balance efficiency and equity.

- When population centers are scattered throughout a region, targeted investments in road improvements have the potential to greatly impact efficiency of health care delivery systems.

Our model aims to provide a reasonably comprehensive description of health care environments encountered in sub-Saharan Africa. However, some of the assumptions may be somewhat limiting. While the assumption that there is an equal and limited number of health workers at each health center reflects one of the challenges of health care in resource-constrained countries, an alternate model that allows for staffing decisions might be of interest. Additionally, the assumption that health workers travel into communities and that communities are always served by the nearest health center ignores alternative models of African health care delivery. A study of the tradeoffs in equity, efficiency, and equity under alternative health care delivery models would also be informative.

While we considered the impact of infrastructure improvements, the costs of such improvements were excluded from the study. An enhanced model in which decisions and costs regarding infrastructure improvements are incorporated might provide additional insights. Another extension of interest is a model that incorporates the preventive value of treatments of communicable diseases.

CHAPTER III

DYNAMIC ASSIGNMENT OF PATIENTS TO HOSPITAL BEDS

3.1 Introduction

A key issue plaguing health care systems both in the U.S. and abroad is a pattern of overcrowding within hospitals, specifically in emergency departments (ED), operating rooms (OR), and intensive care units (ICU). The practice of "boarding" in hospitals is both a cause and effect of overcrowding within hospitals. The boarding of patients refers to a patient continuing to occupy a bed once the patient has been designated for hospital admission or transfer to another unit, often due to a bed being unavailable in the next unit. ED boarding is a particularly serious problem with an estimated 17% of total ED hours between 2003 and 2005 being attributed to boarding patients [13]. These boarding patients cause bottlenecks in the hospital system preventing new patients from being admitted, decreasing the number of people treated, and decreasing hospital revenues [21]. In addition to impact on throughput, the quality of care delivered to patients is negatively impacted by the practice of boarding [39, 62]. In fact, patients who are likely to experience long periods of boarding, such as the elderly, are more likely to have poor health consequences as a result [38].

One of the primary causes of boarding is unavailability of beds downstream from the ED, ICU, and OR departments [33, 67]. Correspondingly, a system-wide perspective is needed to approach the problem of reducing overcrowding and boarding throughout the hospital. Unfortunately, the problem of reducing bottlenecks is complex and requires consideration of many factors. The task of balancing both demand and supply of hospital beds is overseen by a bed manager or bed management team [9].

Due to numerous competing requests for limited hospital resources and unique clinical needs of patients, the task of choosing when and how to make an assignment is often challenging, especially in an overcrowded hospital [67]. By strategically making bed assignment decisions, bed managers are able to alleviate the negative consequences of overcrowding including a detriment to safety, timeliness, and effectiveness, all key components of quality health care as defined by the Institute of Medicine [40].

Due to the complexity of hospital systems, bed managers have multiple options when considering the assignment of a patient to a bed. For instance, they may assign the patient to the ideal bed (if it is available), postpone assignment through boarding, or hold the patient in a separate temporary location, such as a hallway [37]. Naturally, the bed assignment for a particular patient is linked to the nature of his or her condition. For example, a patient recovering from heart surgery may be prioritized to be assigned to a bed in the cardiology wing. In general, expertise by the health care professionals in a particular wing or area of the hospital leads to improved patient outcomes and safety, and more efficient rounding by physicians. However, when a hospital is highly capacitated, a bed in the ideal unit may not be available and instead the patient may need to be placed in an alternate unit with available bed capacity and appropriate resources. This practice is often referred to as overflow or "off-service" assignment and may negatively impact patient care [34, 53]. The impact of a patient's assignment to an overflow unit must be weighed against the similar effects from boarding, resulting in a careful balancing of tradeoffs by the bed manager.

Decision making related to bed assignments is complicated by the dynamic nature of hospitals including the constant change in status and expectations for patient admissions, transfers, and discharges. Historically, in many hospitals, the process of transferring patients between beds was completed through one-on-one communication between hospital staff, in a decentralized manner. As a result, bed assignments

were made with limited information about the system-wide availability of beds and patients' needs, potentially leading to inequity among patients. In recent years, hospitals have increasingly adopted information systems and software packages to aid bed managers by providing visibility i.e., real-time information about the status of patients, rooms, resources, and bed requests [6]. While these systems are useful for visually processing information about the hospital system, to the best of our knowledge they do not provide decision support, e.g., about the impact a particular bed assignment on patient flow and the ability to fulfill bed requests in the future. We explore the use of analytical approaches for real-time dynamic bed assignments in hospitals utilizing information contained in these information systems.

We are interested in the process by which a patient currently occupying a bed in one unit of the hospital (e.g., ED or ICU) is transferred to a bed in another unit of the hospital. First, upon entering the system the patient occupies an ED (or ICU) bed in the upstream unit. When the patient no longer requires care in the ED and is ready to be transferred, the patient is either assigned to a bed in a downstream unit or boards in the current unit (i.e., continues to occupy a bed) while waiting for a bed assignment. The set of downstream units to which a patient can be feasibly assigned is directly related to the nature of the patient's illness, or the type of patient. Therefore, if no downstream beds are available, "blocking" may occur (due to boarding patients) in the ED preventing new patients from being admitted. While waiting for an assignment, it is possible that a patient completes his or her hospital stay and leaves directly from the ED.

We develop a model of patient flow within a hospital and algorithms to address the bed assignment decisions faced by bed managers. In our model we consider multiple patient classes and hospital units. The primary objective is to maximize the appropriateness of assignments of patients to units when both patients and beds become available in an online fashion. The second objective is to minimize the time

patients wait for assignment, or board in an upstream unit. Using a Markov Decision Process (MDP) we model the hospital system as a tandem queueing network with multiple customer classes and cross-trained server pools, develop new algorithms for dynamic patient assignment, and test the performance against current hospital practices through simulation.

We demonstrate that implementation of analytical methods that utilize real-time information about the hospital system, similar to that provided by bed management information systems, can result in the simultaneous improvement of multiple performance metrics. These metrics include the number of patients served, the rate of off-service assignments, mean boarding time, and median boarding time. These improvements are seen in systems with a variety of costs for waiting, average system utilization, and alignment of supply and demand for individual units. These algorithms result in greater improvement when the hospital places a priority on the appropriateness of the assignments (rather than the boarding times and costs). When boarding costs are high, representing the hospital's preference for reducing congestion in upstream units, the improvement from the use of these analytical methods is limited.

In the remainder of this chapter we first provide an overview of relevant literature (Section 3.2). We then present a model of the hospital system and associated bed assignment decisions in Section 3.3. Utilizing the structure of the model, we present our dynamic bed assignment algorithms in Section 3.4. We test the performance of the proposed algorithms with a discrete event simulation and numerical study in Section 3.5. Finally we propose areas for future research and summarize the research contributions in Sections 3.6 and 3.7, respectively.

3.2 Literature Review

The model we develop for the bed assignment problem shares similarities with well-known problems from the operations research literature such as the online assignment or matching problem and the problem of customer routing in queueing systems. In our review of literature, we first examine models that have been developed to study boarding and overflow within hospitals in Section 3.2.1. In Sections 3.2.2 and 3.2.3 we review related literature concerning dynamic assignment problems and queueing models.

3.2.1 Models of Patient Overflow and Boarding

Due to the threat that overcrowding and boarding pose for patient throughput and quality of care, researchers study both the causes of and potential solutions to overcrowding and excessive boarding in hospitals. Levin et al. (2008) develop a discrete event simulation model to study the impact of demand for cardiology beds by other units, including telemetry, on the rate of ED boarding of cardiology patients. They find that reserving beds for potential telemetry patients is a significant cause of the excessive blocking, in which patients are prevented from moving to the next bed due to the lack of available beds. The simulation does not explicitly model the bed assignment process, but rather assumes that patients' movements are similar to historical data [45]. Bretthauer et al. (2011) address the role of boarding in hospitals and develop a heuristic for tandem queues to evaluate the impact of blocking on patient flow [11]. While the algorithm they develop does provide a good approximation for the effect of blocking, the system presented is simplified by the assumptions of homogeneous patients and fixed probabilities for routing patients between units, or general patient routing. While our model also addresses blocking, we also allow for state-dependent routing and heterogeneous patients.

Other researchers have developed models to identify strategies for decreasing the

rates of overflow and boarding. Harrison et al. (2005) develop a discrete event simulation model of patient flow in an acute care hospital to examine the impact of bed capacity on patient overflow. They note that seasonality of arrivals throughout the year is a leading cause of overflow [34]. Another approach to dealing with overcrowding is the use of ED admission control policies. Helm et al. (2011) develop a model of a hospital system in which call-in patients request a bed and do not arrive until a bed is available. In order to decrease overcrowding they consider the placement of a patient in a "non-preferred" unit and seek to evaluate the off-unit census under a variety of admission control policies [36]. While their initial model considers homogeneous patients, their simulation considers multiple patient types [36]. Rather than considering all overflow units to be equivalent, our model addresses the appropriateness of the assigned unit.

While significant work has been done related to admission control policies, there is relatively limited research related to the decisions of when to assign patients to non-preferred units, or allow for overflow. To the best of our knowledge, Teow et al. (2011) are the first to consider the problem of identifying the causes of "overflow beds" through data mining and statistical techniques. They suggest that statistical information can be used to guide the bed management process of determining when to use an overflow bed, when to wait for an appropriate bed, and which units are most appropriate for use of overflow beds [80]. In comparison to existing literature, we address the process of choosing when and which assignments to make, including postponing assignments, and consider the value of the appropriateness of the unit in this decision process.

3.2.2 Dynamic Assignment and Online Matching Problems

The bed manager's decision is to assign available patients to available beds where each feasible assignment has an associated value and both beds and patients become

available in an online fashion. The bed assignment problem is similar to an online bipartite matching, or online assignment problem, in which nodes and corresponding edges become available over time. In an online assignment problem, upon arrival of a node, and corresponding edges, an assignment of two nodes is made, removing them from the graph. A reward, defined by the objective function, is received upon assignment. The standard online assignment problem seeks to maximize or minimize the total weight of the edges corresponding to the assignments. Here we use the terms online bipartite matching and online assignment problems interchangeably.

Algorithms have been developed for many variants of the online assignment problem. Khuller, Mitchell, and Vazirani (1994) develop an online deterministic algorithm for an online stable marriage problem where each node has a preference for nodes from the other set. With one set of nodes arriving one-by-one in an online fashion, each arriving node must immediately be matched. Unlike standard assignment problems, their algorithm seeks to find a stable matching to minimize the dissatisfaction with matches, or the preference to have been in another match [41]. The different preferences by each set of nodes has similarities with the bed assignment problem. While a value or weight is assigned to each pair of patient and bed, it is not guaranteed that the unit is the best match for the patient, and vice versa, because preferences or values for beds, or marriage partners, are not drawn from a metric space. Within literature that develops algorithms for the online assignment problems, the assumption of a metric space for defining edge weights is common [18, 25].

Another related stream of literature is that of online stochastic bipartite matching in which the arrivals are drawn from a known or unknown [47, 82] distribution. Here the objective is to maximize or minimize the sum of the weights associated with the matchings. A common application of online stochastic bipartite matching is the modeling of allocation of internet advertisements [5, 12, 22, 50]. In most online assignment problems, an assignment must be made immediately upon arrival, or

the arrival is discarded without assignment. While this assumption is reasonable in many settings, it corresponds to rejecting a patient which may not be a possibility in hospitals. Instead, we consider the option of postponing the assignment by boarding a patient until he or she is assigned to a bed in an appropriate unit.

Another factor distinguishing the dynamic bed assignment problem from other online assignment problems is the existence of advance information about arrivals of both beds and patients. Advance information can take the form of a "look-ahead" interval, in which information about events in the future interval is (partially) available when a decision is made. Spivey and Powell (2004) study a similar dynamic assignment problem in a freight transportation setting. After providing a general definition for the dynamic assignment problem, Spivey and Powell develop an adaptive algorithm which accounts for the value of advance information, in the form of a "look-ahead" interval. While this work makes contributions in defining the general dynamic assignment problem, in their numerical experiments they assume a constant fixed set of entities from which arrivals occur or a metric space for evaluating the value of an assignment [76]. Their algorithm disregards the reuse of resources, such as trucks or hospital beds. Therefore the availability of a particular resource is not related to the assignment of that resource in the past. With our consideration of hospital beds, the probability a bed becomes available is a direct consequence of assignment decisions from past periods and service times of previously assigned patients. Due to the reuse of resources (beds), there are no direct variations on the algorithms developed by Spivey and Powell that are applicable in our setting. The algorithms we develop assume known probabilities of arrival for both tasks and resources in an assignment process and do not include an option to refuse assignment if capacity is available.

While the problem we consider shares similarities with online assignment problems, different characteristics of the bed assignment problem and model include the

following

- Two streams of online arrivals (patient arrivals and beds becoming available after completion of service)
- No requirement for immediate assignment (a patient may board until an appropriate bed becomes available)
- Arrival patterns of beds are impacted by past assignment decisions, i.e., there is reuse of resources (a bed becomes available when a previously assigned patient completes service)
- Edge weights (the value of a patient-bed assignment) are not drawn from a metric space

As a result, much of the literature related to online and dynamic assignment problems is not directly applicable.

3.2.3 Control of Queueing Systems

The need to consider the interdependence of arrivals, assignments, and the future availability of resources directly relates to a queueing network with competing requests for limited resources. It is common for hospitals to be modeled as a queueing network, in which patients are in service or waiting for beds [48, 52]. A key characteristic distinguishing hospitals from traditional queueing systems is the lack of a physical queueing area between units. This leads to the concept of "blocking" or "boarding", i.e., a patient whose service is completed in one unit continues to "board" in that unit until a bed becomes available in the next unit to which the patient is assigned. Blocking can be represented with an imaginary queue in which blocked customers are waiting for resources to become available. A variety of research examines blocking in queueing systems [10, 27, 77]. As mentioned above, Bretthauer et al. (2011) develop a heuristic to accurately represent the effects of blocking in a hospital system [11].

We consider blocking, cross-trained servers, the need for state-dependent routing policies, and class-dependent service rates in our model. Cross-trained servers refer to the existence of flexible servers which can serve multiple patient types, or can serve multiple functions in the queueing system. Queueing models have been developed that examine routing in systems with multiple customer classes [30], flexible servers [26], and the need for state-dependent policies for control of such systems [51, 60]. Many of these control problems examine the switching of the designation of servers rather than the routing of customers to servers. Development of policies for routing customers has been examined under a variety of settings [48, 54, 55, 60, 75]. In the presence of cross-trained servers, the development of optimal routing policies becomes more difficult. Models with class-dependent service rates have also been developed [35, 63]. While blocking, cross-trained servers, state-dependent routing policies, and class-dependent service rates have been studied independently, we are not aware of research that simultaneously examines all of these features. Additionally, we consider the objective of maximizing the appropriateness of a match between customer and server (through routing policies) rather than only minimizing the average wait time, distinguishing the problem we examine from previous studies of queueing systems.

3.3 Model

We model the system as a tandem queueing network with multiple cross-trained server pools and customer classes, and analyze it as a Markov Decision Process. In the tandem queueing network model, patients, beds, and hospital units are represented as customers, servers, and server pools, respectively. There are I customer classes and $J + 1$ server pools. Pool 0 corresponds to the first stage of service, or the upstream unit. Pools $j \in \{1 \dots J\}$ refer to the parallel server pools in the second stage of service, or downstream units. Pool k has N_k identical servers (beds) for $k \in \{0 \dots J\}$.

In practice, patients may enter the system through many different upstream units.

In focusing our attention on the bed assignment process, we assume the existence of one upstream unit to which all patients are admitted before assignment to one of many downstream units. An extension to consider the assignment of patients with multiple upstream units is left for future work and discussed in Section 3.6.

Customers from class i arrive to the system following a Poisson process with rate λ_i. Servers in the upstream pool 0, are non-idling and any customer who sees an idle server in pool 0 upon arrival immediately begins service. If all upstream beds are occupied an arriving patient is diverted to another unit or hospital, i.e., if all servers in pool 0 are busy, any arriving customers immediately leave the system. Therefore, we do not address queues that may form for beds in the upstream unit. While we acknowledge that admission control policies for the upstream unit do play a role in the boarding and overflow that occurs within a hospital, in order to focus on the role of bed assignments, we assume that patients are admitted in order of arrival.

Service time for customers of class i by a server in pool 0 is exponentially distributed with rate μ_i^0 for $i \in \{1 \dots I\}$. Previous work has confirmed the appropriateness of modeling a hospital system with Poisson arrivals and exponential service times [28, 52, 57]. A customer class can be served by multiple downstream server pools which are cross-trained to serve multiple customer classes. Service time for customers of class i by a server in pool $j \in \{1 \dots J\}$ is exponentially distributed with rate μ_i for $i \in \{1 \dots I\}$. Let E be the set of all pairs of customer classes and downstream server pools such that the customers are eligible for assignment to the unit (Equation 1).

$$E = \{(i,j) : \text{servers in pool } j \text{ can feasibly serve customers in class } i;\ i \in \{1 \dots I\}, j \in \{1 \dots J\}\} \tag{1}$$

The benefits received from the appropriateness of the bed assignment are impacted by the length of time the patient spends in the unit. Therefore, for a pair $(i,j) \in E$, a value $a_{i,j}$ is achieved for each unit of time a customer of class i is in service in pool

j. Sometimes a customer who completes service in pool 0 continues to wait (i.e., "board") there for assignment, for example, if appropriate servers in the downstream unit are all occupied or if some of the available servers are reserved for potential future arrivals. For each unit of boarding time, there is a penalty b_i for a customer in class i. While boarding, a patient's recovery and treatment continues as if he is assigned to a downstream unit. Therefore, a boarding customer of class i leaves the system with rate μ_i. Since we assume that the length of time needed for recovery or treatment is not dependent on the location of the patient, the rate of a customer exiting the system while boarding, or waiting in queue, is the same as the rate at which the customer departs from the downstream server pools.

The system and associated processes can be represented as a tandem queueing system with an imaginary queue existing between upstream and downstream servers. A depiction of the system with two customer classes, one upstream server pool, and three downstream server pools is shown in Figure 13.

Considering both the cost of boarding (b_i) and the value of assignment ($a_{i,j}$) in the system, we are interested in identifying patient-bed assignment policies that minimize the difference between the total value and total cost. Once a customer completes service in the upstream server pool, either he is immediately assigned to one of the servers in the downstream pools, or his assignment is delayed. The delay can be due to the unavailability of a feasible downstream server, or due to the desire to reserve a downstream server for a potential future arrival. Therefore, it is possible to have a compatible idle server and a boarding customer concurrently.

Assuming that assignment decisions are only made at a change to the system state (i.e., patient arrival or service completion) and that the time between state changes is exponentially distributed, the system can be modeled as a continuous-time Markov decision process. Correspondingly, applying the uniformization technique we formulate the equivalent average reward criterion infinite-horizon discrete-time

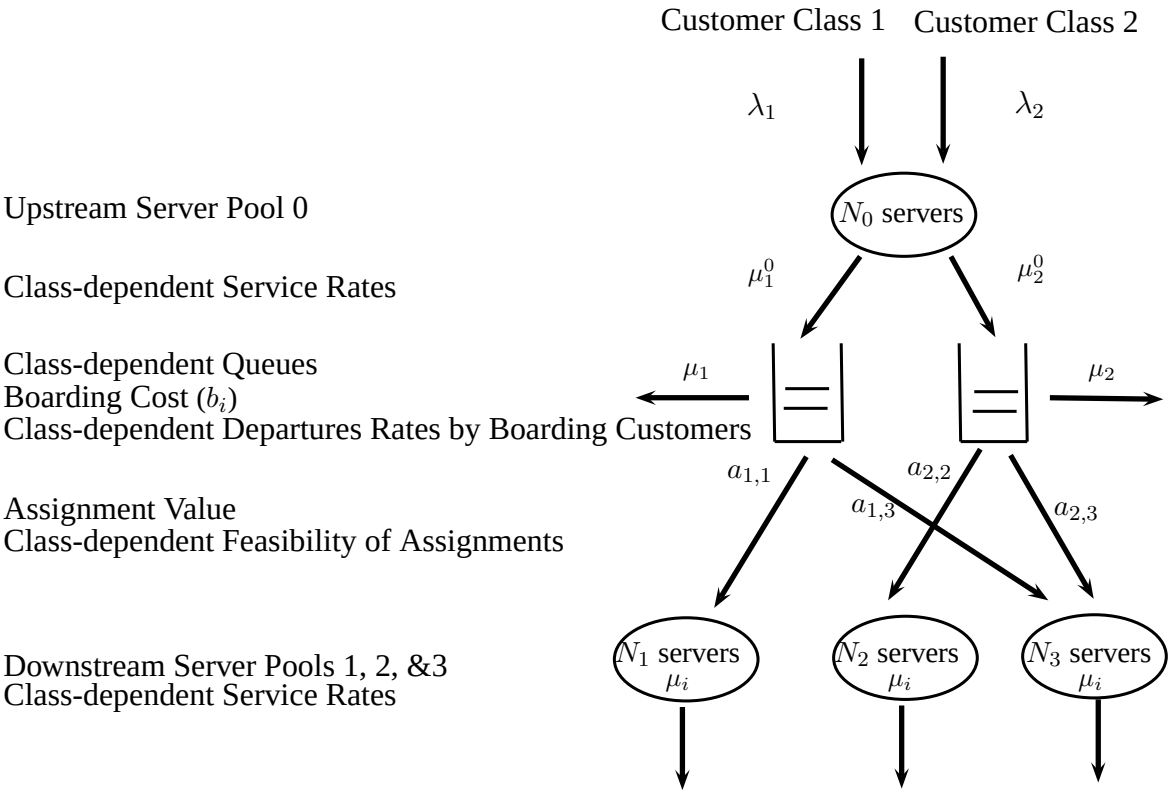


Figure 13: Description of patient assignment processes as a queueing system

Markov Decision Process (MDP) below [2, 68].

3.3.0.1 State Space

The state of the system is defined by the number of customers in service or waiting for assignment in upstream or downstream server pools. A state s is defined as $s = (\vec{U}, \vec{B}, \vec{D})$. Let $\vec{U} = (U_1, U_2, \ldots, U_I)$, $\vec{B} = (B_1, B_2, \ldots, B_I)$, and $\vec{D} = (D_{1,1}, D_{1,2}, \ldots, D_{1,J}, D_{2,1}, \ldots, D_{2,J}, \ldots, D_{I,J})$ where U_i and B_i are the number of upstream servers that are busy due to service or boarding of a customer of class i, respectively. $D_{i,j}$ is the number of servers in pool j that are busy serving a customer of class i.

The state space is restricted such that the total number of customers in service or boarding upstream does not exceed the upstream server pool capacity, N_0. Similarly, the number of customers in service in unit j cannot exceed N_j. Finally, $D_{i,j}$ can only

be positive when an assignment of a customer of class i to server pool j is feasible (i.e., $(i,j) \in E$). The state space, S, is defined in Equation 2.

$$S = \{(\vec{U}, \vec{B}, \vec{D}) | \sum_{i=1}^{I} U_i + B_i \leq N_0; \sum_{i:(i,j)\in E} D_{i,j} \leq N_j \ \ \forall j \in \{1 \dots J\}; D_{i,j} = 0 \ \ \forall (i,j) \notin E\} \tag{2}$$

When in some state s the transition to the next state happens due to one of the following: (i) the arrival of a new customer, (ii) completion of service by a customer in the upstream server pool, (iii) the completion of service of a boarding customer (while waiting for assignment), (iv) the completion of a customer being served by a downstream server or (v) the action of assigning boarding customers to available servers in downstream pools.

The transitions from the stochastic processes (i)-(iv) result in decision epochs. While the arrival of a new customer results in a decision epoch, we assume no action is taken as the set of possible actions does not change. The newly arriving customer is assigned to a server in the upstream unit, while the number of boarding customers and available downstream servers remains the same. Additionally, we assume that after the completion of service by a boarding customer (iii), no action is taken as the set of feasible actions is a subset of the feasible actions prior to the service completion.

There are at most $3I + |E| + 1$ states that may be transitioned to from any state s that do not relate to an action. For instance, there are I states that may result from a customer completing service in the upstream server pool (ii), represented as s_{+B_i} (Equation 3).

$$s_{+B_i} = (U'_1, \dots, U_{i-1}, U_i - 1, U_{i+1}, \dots, U_I, B_1, \dots, B_{i-1}, B_i + 1, B_{i+1}, \dots, B_I, D_{1,1}, \dots, D_{I,J}) \tag{3}$$

Alternatively, the system changes upon completion of service by a boarding customer (iii). The system then transitions to state s_{-B_i} (Equation 4).

$$s_{-B_i} = (U_1, \dots, U_I, B_1, \dots, B_{i-1}, B_i - 1, B_{i+1}, \dots, B_I, D_{1,1}, \dots, D_{I,J}) \tag{4}$$

With the completion of service of a customer of type i in a downstream server pool j, (iv), a new server becomes available in unit j. The definition of the state resulting from service completion in a downstream unit is provided in Equation 5. There are at most $|E|$ states that can result from service completion of a customer in a downstream server pool.

$$s_{-D_{i,j}} = (U_1, \ldots, U_I, B_1, \ldots, B_1, D_{1,1}, \ldots, D_{i,j-1}, D_{i,j} - 1, D_{i,j+1}, \ldots, D_{I,J}) \tag{5}$$

In Equation 6, we define the state resulting from the arrival of a new customer (i), s_{+U_i}.

$$s_{+U_i} = (U_1, \ldots, U_{i-1}, U_i + 1, U_{i+1}, \ldots, U_I, B_1, \ldots, B_1, D_{1,1}, \ldots, D_{I,J}) \tag{6}$$

Let F_s be the set of all feasible states that result due to a customer service completion in an upstream server pool (ii) or a downstream server pool (iv). The definition of F_s is provided in Equation 7.

$$F_s = \{s_{+B_i} \quad \forall i \in \{1 \ldots I\} : U_i > 0\} \bigcup \{s_{-D_{i,j}} \quad \forall (i,j) \in E : D_{i,j} > 0\} \tag{7}$$

Let $s = (\vec{U}, \vec{B}, \vec{D})$ be the state immediately after (ii) or (iv). The manager takes an action in state s, which may include the possibility of taking no action. Let $\vec{x}^s = (x^s_{1,1}, x^s_{1,2}, \ldots, x^s_{1,J}, x^s_{2,1}, \ldots, x^s_{I,J})$ represent the action, or assignments, taken in state s. $\vec{x}^s$ must satisfy Equations 8 and 9. The actions are limited by the number of boarding patients and the availability of beds in the downstream unit. The set of feasible actions that can be taken in state s is defined by the set A_s.

$$A_s = \{\vec{x}^s : \sum_{j:(i,j)\in E} x^s_{i,j} \leq B_i \quad \forall i \in \{1 \ldots I\}; \tag{8}$$

$$\sum_{i:(i,j)\in E} x^s_{i,j} \leq N_j - \sum_{i:(i,j)\in E} D_{i,j} \quad \forall j \in \{1 \ldots J\}; \tag{9}$$

$$s = (\vec{U}, \vec{B}, \vec{D})\} \tag{10}$$

Because interarrival times and service times are exponentially distributed, we can model the continuous-time system as a discrete time MDP with the uniformization approach [2, 46, 68, 74]. In this fashion we define the corresponding optimality equations, for the problem of maximizing the average reward over an infinite horizon. We define the uniformization constant β to ensure that it is greater than the transition rate for all possible states $s \in S$. The largest possible transition rate corresponds to all upstream and downstream servers occupied by customers with the greatest possible service rates (Equation 11).

$$\beta = \sum_{i=1}^{I} \lambda_i + \max_i(\mu_i^0, \mu_i) N_0 + \sum_{j=1}^{J} \max_{i:(i,j)\in E}(\mu_i) N_j \tag{11}$$

Let v^* be the optimal expected average reward per unit time. $g^*(s)$ is the value of starting in state s under an optimal policy. Let $\{s\}^{\vec{x}^s}$ be the state resulting from taking action $\vec{x}^s$ in state s.

Upon assignment of a customer to a downstream server pool, a one time reward is received accounting for the expected time the patient spends in the downstream unit, $\frac{1}{\mu_m}$, and the per unit time value of assignment $a_{m,n}$. Due to the memoryless property of the exponential distribution, the expected time a customer spends in the system after assignment is independent of the time spent boarding. Therefore, upon assignment of a customer of type m to server pool n a one time reward of $\frac{a_{m,n}}{\mu_m}$ is received.

The set of Bellman equations for this problem is expressed in Equation 12. The first term in the equation represents the cost of boarding incurred until the next transition. The second term corresponds to the completion of service by a customer in the upstream server pool, incorporating the reward from the assignment $\vec{x}^{s+B_i}$. The third term corresponds to the service completion of a boarding patient in the upstream server pool. The fourth term corresponds to the service completion of a customer in

a downstream server pool, resulting in the availability of a new downstream server. A reward is received for the assignment $\vec{x}^{s_{-D_{i,j}}}$.

A transition to a state with an additional customer in upstream service is only feasible if the upstream pool is not full. Hence, the fifth term corresponds to the arrival of a customer to the system and beginning service. When the upstream server pool is full, upon arrival of a new customer, the system remains in the same state. Additionally, note that with uniformization there is a positive probability of no state transition, corresponding to the final term in the Bellman equation.

$$\frac{v^*}{\beta} + g^*(s) = \max_{\substack{\vec{x}^{s'} \in A_{s'} \\ \forall s' \in F_s}} \begin{cases} \frac{-\sum_{i=1}^{I} B_i b_i}{\beta} + \sum_{i=1}^{I} \frac{\mu_i^0 U_i}{\beta} \left(\sum_{(m,n)\in E} x_{m,n}^{s_{+B_i}} \frac{a_{m,n}}{\mu_m} + g^*(\{s_{+B_i}\}^{\vec{x}^{s_{+B_i}}}) \right) \\ + \sum_{i=1}^{I} \frac{\mu_i B_i}{\beta} g^*(s_{-B_i}) + \sum_{(i,j)\in E} \frac{\mu_i D_{i,j}}{\beta} \left(\sum_{(m,n)\in E} x_{m,n}^{s_{-D_{i,j}}} \frac{a_{m,n}}{\mu_m} + g^*(\{s_{-D_{i,j}}\}^{\vec{x}^{s_{-D_{i,j}}}}) \right) \\ + \sum_{i=1}^{I} \frac{\lambda_i}{\beta} g^*(s_{+U_i}) + \frac{\beta - \sum_{i=1}^{I} (\mu_i^0 U_i + \mu_i B_i + \lambda_i + \sum_{j:(i,j)\in E} \mu_i D_{i,j})}{\beta} g^*(s) \\ \text{if } \sum_{i=1}^{i} (U_i + B_i) < N_0 \\ \frac{-\sum_{i=1}^{I} B_i b_i}{\beta} + \sum_{i=1}^{I} \frac{\mu_i^0 U_i}{\beta} \left(\sum_{(m,n)\in E} x_{m,n}^{s_{+B_i}} \frac{a_{m,n}}{\mu_m} + g^*(\{s_{+B_i}\}^{\vec{x}^{s_{+B_i}}}) \right) \\ + \sum_{i=1}^{I} \frac{\mu_i B_i}{\beta} g^*(s_{-B_i}) + \sum_{(i,j)\in E} \frac{\mu_i D_{i,j}}{\beta} \left(\sum_{(m,n)\in E} x_{m,n}^{s_{-D_{i,j}}} \frac{a_{m,n}}{\mu_m} + g^*(\{s_{-D_{i,j}}\}^{\vec{x}^{s_{-D_{i,j}}}}) \right) \\ + \frac{\beta - \sum_{i=1}^{I} (\mu_i^0 U_i + \mu_i B_i + \sum_{j:(i,j)\in E} \mu_i D_{i,j})}{\beta} g^*(s) \\ \text{if } \sum_{i=1}^{i} (U_i + B_i) = N_0 \end{cases} \tag{12}$$

Due to the large dimension of both the state and action spaces, identifying the exact optimal solution to the MDP is difficult for real-world sized problems with currently available commercial solvers. For example, a hospital with 30 servers in the upstream pool, 15 servers in each of 10 downstream pools, and with customers from one of 10 customer classes which can be feasibly assigned to two downstream

server pools results in 1.75×10^{34} feasible states. By comparison, for a small example, with 5 servers in the upstream pool, 5 servers in each of 2 downstream pools, and with customers able to be feasible assigned to either downstream server pool, 225,225 feasible states would result.

3.4 Bed Assignment Algorithms

With the goal of developing efficient and effective bed assignment policies, we utilize the general structure, notation, and transition probabilities corresponding to the MDP and tandem queueing system to define five dynamic assignment algorithms, GREEDY, VALONLY, PROB, STOCH, and HYBRID. Each algorithm is designed to be applied at each decision epoch, utilizing information about the system state, $(\vec{U}, \vec{B}, \vec{D})$, to identify a bed assignment action.

Each of the algorithms focus on choosing an action by weighing the one period nonassignment costs and the expected value of the state at the next decision epoch. Because the algorithms focus on two stages of decision making, we examine the set of transitions available in the state corresponding to the first period (s) for the definition of β. In the MDP defined above, this corresponds to assigning β to be equal to the transition rate in state s, β_s. Thus, $\beta = \beta_s = \sum_{i=1}^{I} (\mu_i^0 U_i + \mu_i B_i + \lambda_i + \sum_{j:(i,j)\in E} \mu_i D_{i,j})$.

In Section 3.4.1, we define two myopic algorithms, GREEDY and VALONLY, that simulate the decision processes currently used at many hospitals. Next, in Section 3.4.2 we utilize the probabilities of arrivals and departures used in the definition of the MDP and queueing models, to define the PROB algorithm. In Section 3.4.3 we develop the STOCH algorithm which utilizes a stochastic optimization model to identify assignment decisions at each epoch. In Section 3.4.4 we combine features of the PROB and STOCH algorithms to develop the HYBRID algorithm. Finally, we develop the algorithm UB to calculate an upper bound on the total reward in Section 3.4.5.

3.4.1 Myopic Algorithms

To develop a benchmark for comparison, we define two myopic assignment algorithms to approximate the assignment processes currently used by hospital bed management staff. In both algorithms, we assume nonidling downstream beds, such that no bed will remain idle if it can be feasibly assigned to a boarding patient. As a result of this nonidling assumption, at most one assignment will be made at each decision epoch. Both myopic approaches are non-anticipatory, only considering patients and beds that are currently available for assignment and disregarding expectations about future arrivals.

In the first algorithm (GREEDY), we assume that patients are prioritized for assignment based on their length of stay in the hospital. Under this approach, the patient who has been in the hospital system the longest is assigned to the best available and feasible bed. If no beds are available that meet clinical requirements, the patient continues to board in the upstream unit until the next decision epoch. After a patient has been assigned to a bed, the patient with the next highest priority is considered for assignment. This greedy approach seeks to find the best assignment for each patient individually ignoring the overall system performance.

In another myopic algorithm, VALONLY, the feasible assignment with the highest $a_{i,j}$ value is made assigning a boarding patient of type i to unit j. While similar, VALONLY seeks to make the best possible assignment from a systems perspective, while GREEDY prioritizes the patients that have been in the system for the longest time.

3.4.2 Probabilistic Approximation for Non-Assignment of Patients or Beds

In addition to the value from a particular patient-bed assignment there is also an associated opportunity cost. For example, rather than assigning patient i to a bed in unit j and receiving the reward $a_{i,j}$ now, it may be preferable to save the capacity

in unit j for a potential future arrival with a higher reward, or with the goal of avoiding a future off-service assignment. As a result, there is value associated with both assignment and nonassignment.

In the PROB algorithm, we develop an approximation of the value of assigning each patient type to a feasible unit, as well as values of nonassignment for patients and beds. While we have a similar approximation of the assignment and nonassignment values as Spivey and Powell (2004), due to their assumption of no reuse of resources (beds), their adaptive algorithm is not directly applicable in this system. In PROB, we develop approximations by incorporating our knowledge of the probabilities of future events, as used in the definition of the MDP and queueing models. At each decision epoch, we solve an assignment problem which incorporates these assignment and non-assignment values to inform the decisions.

3.4.2.1 Value Approximations

We estimate the value associated both with assignment and nonassignment. We let the pairs $(i, 0)$ and $(0, j)$ to represent the nonassignment of a patient of type i and the nonassignment of a bed in unit j, respectively. $\bar{E}$ is the union of the set of pairs (i, j) contained in E and the pairs associated with nonassignment of patients and beds, $(i, 0)$ and $(0, j)$.

$$\bar{E} = \{(i,j) \in E\} \bigcup \{(i,0) \forall i \in \{1 \dots I\}\} \bigcup \{(0,j) \forall j \in \{1 \dots J\}\} \tag{13}$$

The (short term) value of assigning a patient of type i to a bed in unit j is denoted by $V_{i,j}$. This value represents the expected reward achieved over two periods as a result of the assignment. A reward of $a_{i,j}$ is received for the assignment at t. Unless the assigned patient exits the system at $t+1$, a reward of $a_{i,j}$ is also achieved at decision epoch $t+1$ with no boarding cost (b_i) incurred. The probability of the assigned patient exiting the system at the next decision epoch is $\frac{\mu_i}{\beta_s}$. We define $V_{i,j}$

in Equation 14.

$$V_{i,j} = (2 - \frac{\mu_i}{\beta_s}) a_{i,j} \qquad \forall (i,j) \in E \tag{14}$$

Additionally, for each patient type i we approximate the value of nonassignment of a patient of class i with $V_{i,0}$. With nonassignment of a patient, a cost of b_i, is incurred from the boarding of the patient in the first period t. With probability $\frac{\mu_i}{\beta_s}$ the patient departs the system before the next decision epoch $t+1$, and no reward or cost is incurred for the second period. The patient continues to board in the second period, with a cost of b_i unless a new bed becomes available before the next decision epoch, $t+1$. Here we only consider the arrival of beds that are occupied at t and not the rearrival of beds that were assigned at t. Therefore, if a bed in a compatible unit n becomes available at $t+1$, with probability $\frac{\sum_{m:(m,n)\in E} \mu_m D_{m,n}}{\beta_s}$ a reward of $a_{i,n}$ is received and no boarding cost is incurred in the second period. With these assumptions, we define the approximation for nonassignment of patients (Equation 15).

$$V_{i,0} = -2b_i + \frac{\sum_{(m,n)\in E} \mu_m D_{m,n}(a_{i,n} + b_i)}{\beta_s} + \frac{\mu_i b_i}{\beta_s} \qquad \forall i \in \{1 \dots I\} \tag{15}$$

Similarly, for each bed in unit j we approximate the value of nonassignment of a bed in unit j with $V_{0,j}$. While, no cost is incurred for the nonassignment of a bed, with the decision to defer assignment, a reward may be received with the arrival of a new patient before time $t+1$. The probability of the arrival of a patient of type i is $\frac{\mu_i^0 U_i}{\beta_s}$. We define the expected value of nonassignment of a bed in unit j in Equation 16.

$$V_{0,j} = \sum_{i:(i,j)\in E} \frac{\mu_i^0 U_i a_{i,j}}{\beta_s} \qquad \forall j \in \{1 \dots J\} \tag{16}$$

Utilizing the values of assignment and nonassignment defined above, in the PROB algorithm, at each decision epoch the assignment problem (Equation 17) is solved and the assignments corresponding to the optimal solution are implemented.

$$
\begin{aligned}
\textbf{max} \quad & \sum_{(i,j)\in\bar{E}} V_{i,j}x_{i,j} & & \\
\text{s.t.} \quad & \sum_{j:(i,j)\in\bar{E}} x_{i,j} = B_i & & \forall i \in \{1\dots I\} \\
& \sum_{i:(i,j)\in\bar{E}} x_{i,j} \leq N_j - \sum_{i:(i,j)\in\bar{E}} D_{i,j} & & \forall j \in \{1\dots J\} \\
& x_{i,j} \geq 0 & & \forall (i,j) \in \bar{E} \\
& x_{i,j} \text{ integer} & & \forall (i,j) \in \bar{E} \qquad (17)
\end{aligned}
$$

This algorithm assumes that the value of not assigning patients of type i is linear with respect to the number of patients. As shown by Spivey and Powell (2004), the value obtained from two similar resources or customers is sub-additive. This sub-additive property accounts for competition among patients for beds and is not represented in this algorithm.

3.4.3 Two-Stage Stochastic Program with Recourse (STOCH) Algorithm

Similar to PROB, STOCH seeks to optimize the sum of the expected value in periods t and $t+1$. Unlike PROB, STOCH considers the competition of patients for beds. In STOCH, we formulate a two-stage stochastic program with recourse model to represent the two stages of decision making. The first stage corresponds to the assignment decision made at t. The second stage corresponds to the second decision epoch $(t+1)$. $x_{i,j}$ corresponds to the first stage assignment decisions. Independent of the first stage assignments, a realization of a change in the system state occurs at $t+1$. The change to the system state can include the arrival of a new assignment request by a patient of type i or the new availability of a bed in unit j, resulting in a total of $K = I + J$ scenarios. Unlike PROB, in which we incorporated the probability

of a boarding patient exiting before assignment, we do not incorporate the possibility of a patient leaving from boarding in STOCH, as nonlinear constraints would be needed to account for these scenarios in which the probability of the scenario would be a function of the first stage assignment decisions.

In response to the first stage decisions and realization of the event k, a recourse assignment $y^k_{i,j}$ is made in the second stage. The objective of the two-stage stochastic problem is to identify a robust first stage decision to optimize the expected value achieved after the realization and recourse. Based on the notation and transition probabilities in the definition of the MDP, we define (Equations 18 and 19) and estimate the probabilities of each event k, p_k (Equation 20). Let $\Theta^1_{i,k}$ and $\Theta^2_{j,k}$ be the realizations of uncertainty in scenario k. $\Theta^1_{i,k}$ corresponds to an additional patient of type i becoming available in decision epoch $t+1$ in scenario k. Similarly, $\Theta^2_{j,k}$ corresponds to an additional bed in unit j becoming available in decision epoch $t+1$ in scenario k.

$$\Theta^1_{i,k} = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{otherwise} \end{cases} \qquad \forall i \in \{1 \dots I\}, k \in \{1 \dots K\} \tag{18}$$

$$\Theta^2_{j,k} = \begin{cases} 1 & \text{if } k = I + j \\ 0 & \text{otherwise} \end{cases} \qquad \forall j \in \{1 \dots J\}, k \in \{1 \dots K\} \tag{19}$$

$$p_k = \begin{cases} \dfrac{\mu^0_k U_k}{\sum\limits_{i \in \{1 \dots I\}} \mu^0_i U_i + \sum\limits_{(i,j) \in E} \mu_i D_{i,j}} & \text{if } k \leq I \\ \dfrac{\sum\limits_{(i,k-I) \in E} \mu_i D_{i,k-I}}{\sum\limits_{i \in \{1 \dots I\}} \mu^0_i U_i + \sum\limits_{(i,j) \in E} \mu_i D_{i,j}} & \text{otherwise} \end{cases} \qquad \forall k \in \{1 \dots K\} \tag{20}$$

Since we disregard the probability of patients unassigned at t leaving the system, an assignment in the first period achieves a reward of $2(a_{i,j} + b_i)$, representing the value for two periods. Note that because every patient is either assigned or continues

to board maximizing the function $\sum_{(i,j)\in E}(a_{i,j}+b_i)x_{i,j}$ is equivalent to maximizing $\sum_{(i,j)\in E} a_{i,j}(D_{i,j}+x_{i,j}) - \sum_{i\in 1\ldots I} b_i(B_i - \sum_{(i,j)\in E} x_{i,j})$. For assignments made in the second stage, $y^k_{i,j}$, the reward of $a_{i,j}+b_i$ is achieved for the second period. The full formulation of the two-stage stochastic program with recourse is defined below (Equation 21).

$$
\begin{aligned}
\textbf{max} \quad & \sum_{(i,j)\in E} 2(a_{i,j}+b_i)x_{i,j} + \sum_{k\in\{1\ldots K\}} p_k \sum_{(i,j)\in E}(a_{i,j}+b_i)y^k_{i,j} & \\
\text{s.t.} \quad & \sum_{j:(i,j)\in E} x_{i,j} \le B_i & \forall i \in I \\
& \sum_{i:(i,j)\in E} x_{i,j} \le \sum_{(i,j)\in E} D_{i,j} & \forall j \in J \\
& \sum_{j:(i,j)\in E} x_{i,j} + \sum_{j:(i,j)\in E} y^k_{i,j} \le B_i + \Theta^1_{i,k} & \forall i \in \{1\ldots I\}, k \in \{1\ldots K\} \\
& \sum_{i:(i,j)\in E} x_{i,j} + \sum_{i:(i,j)\in E} y^k_{i,j} \le \sum_{i:(i,j)\in E} D_{i,j} + \Theta^2_{j,k} & \forall j \in \{1\ldots J\}, k \in \{1\ldots K\} \\
& x_{i,j} \ge 0 & \forall (i,j) \in E \\
& x_{i,j} \text{ integer} & \forall (i,j) \in E \\
& y^k_{i,j} \ge 0 & \forall (i,j) \in E, k \in \{1\ldots K\} \\
& y^k_{i,j} \text{ integer} & \forall (i,j) \in E, k \in \{1\ldots K\}
\end{aligned}
\tag{21}
$$

The algorithm STOCH solves this two-stage stochastic program with recourse problem at each decision epoch and implements the assignment defined by the optimal first-stage decision variables, x. While patients exiting the system from boarding cannot be accounted for, this approach does account for the competition among patients for limited beds and the sub-additive properties of postponing assignment of patients or beds.

3.4.4 Combination of STOCH and PROB Algorithms

The STOCH algorithm does not account for patients leaving from boarding while the PROB algorithm does not account for the competition among patients for limited beds. To develop an algorithm that incorporates both of these characteristics, we develop a HYBRID algorithm which has the structure of the STOCH algorithm (Equation 21) but restricts assignments to those that would be in an optimal solution to the problem in Equation 17.

A necessary condition for all optimal solutions to the problem in Equation 17 is that $x_{i,j}$ is only be positive if the sum of the value from nonassignment of the patient ($V_{i,0}$) and nonassignment of a bed in the unit ($V_{0,j}$) is greater than the value from assignment ($V_{i,j}$). Thus we define $\bar{\bar{E}}$ to be the set of pairs (i, j) meeting this criteria (Equation 22).

$$\bar{\bar{E}} = \{(i,j) : (i,j) \in E; V_{i,j} \geq V_{i,0} + V_{0,j}\} \tag{22}$$

Utilizing this restricted set of feasible pairs (i, j), we define the HYBRID algorithm which solves a variant of the integer program in Equation 21 such that the set of decision variables is restricted to those defined by $\bar{\bar{E}}$ rather than E.

3.4.5 Upper Bound on Algorithm Performance

In comparing the performance of the assignment algorithms, we are interested both in the improvement with respect to the myopic approaches as well as the performance with respect to the optimal assignment decisions. Even with perfect information of all arrival and service times for a specific instance, solving the associated integer program to identify the optimal solution is difficult for large instances. The complexity in solving the problem stems from the need to consider reuse of resources, guarantee that capacity restrictions are met at each decision epoch, and ensure nonidling upstream servers. The integer program which incorporates complete information about patient arrivals and service times is provided in Appendix B. With this instance, a small

instance with 100 patient arrivals, 10 downstream units, and each patient able to be feasible assigned to 3 units requires 180,000 integer decision variables and 273,000 constraints. A similar instance with 1,000 patient arrivals requires 18 million integer variables and 27 million constraints.

We utilize the UB algorithm to calculate an upper bound on the optimal reward when it is guaranteed that no patients are diverted for all optimal assignments. We can create a system guaranteeing that no patients are diverted with the optimal decisions by augmenting the instance and increasing the capacity of the upstream unit to a high enough level.

Similar to the other algorithms we have developed, the UB algorithm solves an integer program at each change to the system state. The key difference is the relaxation of the assumption that once a patient is assigned to a unit he remains in the unit until he exits the system. This UB algorithm instead allows for the reassignment of patients at each decision epoch. This includes the possibility of an assigned patient being reassigned to board in the upstream unit or reassigned to a different downstream unit. Let the set $\hat{E}$ be the union of the set E and all pairs (i,0) representing assignment of a patient to the upstream unit. At each decision epoch, all patients not in service upstream are assigned to an upstream or downstream unit according to the solution of the integer program defined in Equation 23, which optimizes the total reward achieved in the associated period. Note, that while we allow for assignment of patients to the upstream unit, we only consider the value obtained for assignments to feasible units (E).

$$
\begin{aligned}
\textbf{max} \sum_{(i,j)\in E} (a_{i,j} + b_i)x_{i,j} & \\
s.t. \sum_{j:(i,j)\in \hat{E}} x_{i,j} = \sum_{j:(i,j)\in E} D_{i,j} + B_i & \qquad \forall i \in \{1 \dots I\} \\
\sum_{i:(i,j)\in \hat{E}} x_{i,j} \leq N_j & \qquad \forall j \in \{0 \dots J\} \\
x_{i,j} \geq 0 & \qquad \forall (i,j) \in \hat{E} \\
x_{i,j} \text{ integer} & \qquad \forall (i,j) \in \hat{E} \qquad (23)
\end{aligned}
$$

As is shown in Theorem 3.4.1, the total reward achieved from the UB algorithm is an upper bound on the optimal objective value when no patients are diverted from the system under either policy. Unfortunately the assumption that no patients are diverted in the optimal solution is necessary to ensure an upper bound, as we have identified instances for which UB performs worse than other algorithms without this assumption.

Theorem 3.4.1. *Assume we have the complete information of all arrivals and service times for a particular instance and let V^* be the total reward achieved when the optimal action is taken at each decision epoch. Similarly, let V^{UB} be the total reward achieved when the action defined by the UB algorithm is taken at each decision epoch. If no patients are diverted from the system when optimal actions are taken at each decision epoch or when using the UB algorithm to choose actions, $V^{UB} \geq V^*$.*

Proof. 3.4.1 Let $s_t^* = (\vec{U}^*(t), \vec{B}^*(t), \vec{D}^*(t))$ be the state of the system at time t when the optimal action is taken at each decision epoch accounting for full information about future events. Additionally, let $\vec{x_t}^*$ be the optimal action taken at decision epoch t. Since optimal actions are made with complete information about future events, we assume that the choice of action is a function of the time t rather than the state s_t^*. Thus, the reward received under the optimal strategy at decision epoch t is

$r(s_t^*, \vec{x_t}^*) = \sum_{(i,j)\in E} a_{i,j}(D_{i,j}^*(t) + x_{i,j,t}^*) - \sum_{i\in\{1\ldots I\}} b_i(B_i^*(t) - \sum_{(i,j)\in E} x_{i,j,t}^*)$. Correspondingly, $V^* = \sum_{t\in T} r(s_t^*, \vec{x_t}^*)$.

Let $s_t^{UB} = (\vec{U}^{UB}(t), \vec{B}^{UB}(t), \vec{D}^{UB}(t))$ be the state of the system at decision epoch t when the UB algorithm is used to identify the action at each decision epoch. Since all patients can be feasibly reassigned at each decision epoch, the feasible action space at t is $A_{s_t^{UB}}^{UB}$ as defined in Equation 26.

$$A_{s_t^{UB}}^{UB} = \{\vec{x_t} : \sum_{j:(i,j)\in\hat{E}} x_{i,j,t} = \sum_{j:(i,j)\in E} D_{i,j}^{UB}(t) + B_i^{UB}(t) \qquad \forall i \in \{1\ldots I\}; \tag{24}$$

$$\sum_{i:(i,j)\in\hat{E}} x_{i,j,t} \leq N_j \qquad \forall j \in \{0\ldots J\}; \tag{25}$$

$$\vec{x_t} \text{ integer} \qquad \} \tag{26}$$

Let $x_t^{\vec{UB}} \in A_{s_t^{UB}}^{UB}$ be the action taken at decision epoch t as defined by the solution to the integer program associated with the UB algorithm (Equation 23). Since all patients may be reassigned at each decision epoch, the states at previous decision epochs do not impact the action space. With these assumptions, the reward at time t is $r(s_t^{UB}, x_t^{\vec{UB}}) = \sum_{(i,j)\in E} a_{i,j}x_{i,j,t}^{UB} - \sum_{i\in\{1\ldots I\}} b_i x_{i,0,t}^{UB}$.

Since no patients are diverted from the system with use of the optimal strategy or the UB algorithm and the assignment decisions do not impact when a patient is in the system, the relationship in Equation 27 holds.

$$B_i^*(t) + \sum_{(i,j)\in E} D_{i,j}^*(t) = B_i^{UB}(t) + \sum_{(i,j)\in E} D_{i,j}^{UB}(t) \qquad \forall i \in \{1\ldots I\} \tag{27}$$

Let us define the action $\vec{z} = (z_{1,0}\ldots, z_{1,J}, z_{2,0}, \ldots, z_{I,J})$ according to Equations 28 and 29.

$$z_{i,j} = D_{i,j}^*(t) + x_{i,j,t}^* \qquad \forall (i,j) \in E \tag{28}$$

$$z_{i,0} = B_i^*(t) - \sum_{j:(i,j)\in E} x_{i,j,t}^* \qquad \forall i \in \{1\ldots I\} \tag{29}$$

From the definition of $\vec{z}$, it is easily seen that $\vec{z} \in A^{UB}_{s_t^{UB}}$. The relationships in Equations 30 - 34 follow.

$$r(s_t^{UB}, \vec{x_t^{UB}}) \geq r(s_t^{UB}, \vec{z}) \tag{30}$$

$$\begin{aligned} r(s_t^{UB}, \vec{z}) &= \sum_{(i,j)\in E} a_{i,j} z_{i,j} - \sum_{i\in\{1\ldots I\}} b_i z_{i,0} \\ &= \sum_{(i,j)\in E} a_{i,j}(D^*_{i,j}(t) + x^*_{i,j,t}) - \sum_{i\in\{1\ldots I\}} b_i(B^*_i(t) - \sum_{(i,j)\in E} x^*_{i,j,t}) \\ &= r(s_t^*, \vec{x_t^*}) \end{aligned} \tag{31}$$

$$r(s_t^{UB}, \vec{x_t^{UB}}) \geq r(s_t^*, \vec{x_t^*}) \tag{32}$$

$$\sum_{t\in T} r(s_t^{UB}, \vec{x_t^{UB}}) \geq \sum_{t\in T} r(s_t^*, \vec{x_t^*}) \tag{33}$$

$$V^{UB} \geq V^* \tag{34}$$

□

As shown in Theorem 3.4.2, in addition to providing an upper bound on the optimal total reward, the UB algorithm provides a tight upper bound on the optimal total reward.

Theorem 3.4.2. *There exists a bed assignment instance such that* $V^{UB} = V^*$.

Proof. 3.4.2 Let there be a bed assignment system with $I = J$ such that for patients in class i the only feasible downstream unit is unit j. Let all $a_{i,i}$ and b_i be positive. Assume that the upstream unit is large enough such that no patients are diverted from the system under either assignment policy.

With these assumptions, this system is equivalent to a system of I parallel multi-server queues with no cross-trained servers. Due to the independence of the queues, the optimal control policy for all is to have non-idling servers in each pool. Therefore, for all instances having this system structure, we define the optimal action at

each time t. If there are $q_i(t)$ patients of class i in the system at time t, the optimal policy is to assign $\min(q_i(t), N_i)$ to server pool i. This results in a reward of $r(s_t^*, \vec{x_t^*}) = \sum_{i \in \{1...I\}} a_{i,i}\min(q_i(t),N_i) - b_i\max(q_i(t) - N_i,0)$ at each epoch t under the optimal assignment policy.

We show that $r(s_t^{UB}, \vec{x_t^{UB}}) = r(s_t^*, \vec{x_t^*})$ for every decision epoch t in any instances corresponding to this system. Assume to the contrary that for some instance and time t, $r(s_t^{UB}, \vec{x_t^{UB}}) > r(s_t^*, \vec{x_t^*})$. Then there exists some t such that $r(s_t^{UB}, \vec{x_t^{UB}}) > \sum_{i \in \{1...I\}} a_{i,i}\min(q_i(t),N_i) - b_i\max(q_i(t) - N_i,0)$. Since the optimal solution to the IP in Equation 23 always assigns $x_{i,i} = \min\ (q_i, N_i)$, we see that $r(s_t^{UB}, \vec{x_t^{UB}}) \leq r(s_t^*, \vec{x_t^*})$ for all t.

Since we have shown that $r(s_t^{UB}, \vec{x_t^{UB}}) \geq r(s_t^*, \vec{x_t^*})$ in Theorem 3.4.1, it follows that $r(s_t^{UB}, \vec{x_t^{UB}}) = r(s_t^*, \vec{x_t^*})$ and $V^{UB} = V^*$ for all instances resulting from this system. □

3.5 Computational Study

To evaluate the relative performance of the algorithms described above (GREEDY, VALONLY, PROB, STOCH, HYBRID) we conduct a numerical study using a discrete event simulation. In this study we address the following research questions.

- What is the impact of new assignment policies on performance metrics such as
  - (i) total reward from the appropriateness of assignment,
  - (ii) mean assignment time,
  - (iii) median assignment time,
  - (iv) rate of off-service assignment
  - (v) the number of patients diverted/accepted?
- How is the benefit from implementation of these algorithms affected by

– (i) the system utilization,

– (ii) balance of supply and demand for specific units,

– (iii) relative weighting of assignment rewards and boarding costs?

The structure of the discrete event simulation and the performance metrics are described in Section 3.5.1. Utilizing the simulation, we conduct comparative experiments using data collected from Children's Healthcare of Atlanta (Children's) for the development of model parameters. Methodology for parameter estimation is provided in Section 3.5.2.

In Section 3.5.3 utilizing the upper bound calculations we examine how use of the three algorithms reduces the optimality gap in comparison with the myopic approaches for "augmented" instance sets, in which the capacity of the upstream unit is increased to ensure no patient diversions. In Section 3.5.4, we expand our analysis to evaluate the relative performance of the algorithms in the original, not augmented, instances. We demonstrate how the difference in performance is impacted by system characteristics. Finally, we evaluate the sensitivity of algorithm performance to changes in b_i and $a_{i,j}$, in Sections 3.5.5 and 3.5.6.

3.5.1 Simulation

We create randomly generated instances which simulate the arrivals and service times of patients over a one year period. Table 6 summarizes the parameter values used in generating the instances along with additional parameters defining the value of assignments ($a_{i,j}$), costs of boarding (b_i), and unit capacities (N_j), we simulate the flow of patients under each policy and evaluate the performance.

With the simulation we calculate the primary performance metrics included in the problem objective, value from appropriateness and cost from boarding (Table 5). Additionally, we examine the secondary performance metrics such as the rate of diversion, rate of off-service assignment, and median assignment times. National

Table 5: Definition of simulation performance metrics

Performance Metric	Description
Assignment Value	Total value achieved from the appropriateness of the assignment of patients to units
Boarding Cost	Total cost incurred from boarding patients in the upstream unit
Total Reward	Difference of assignment value and boarding cost
Mean Assignment Time	The mean time (non-diverted) patients spends boarding
Median Assignment Time	The median time (non-diverted) patients spends boarding
On-Service Assignments	Count of (non-diverted) patients that are assigned to a preferred unit
Off-Service Assignments	Count of (non-diverted) patients that are assigned to a non-preferred unit
Upstream Departures	Count of (non-diverted) patients that are not assigned to a downstream unit
Diversions	Count of the patients diverted from system due to lack of capacity at upstream unit
Accepted	Count of patients not diverted from the system

benchmarks for bed management track the median request-to-assignment time. We hypothesize that some of the delay in real hospital systems is due to manual processes and communication between hospital staff. The median assignment time reported here does not capture the delays due to manual processes, but instead only captures delays due to decisions to postpone assignment or due to the lack of available beds. As such, the median assignment times are significantly lower than what is reported in practice and this difference demonstrates the value of automated bed assignment decision support tools. A description of each of the performance metrics is provided in Table 5.

To remove bias in the calculation of performance metrics we designate a warm-up period of 15 days in the simulation to disregard data collection when the system is not in a steady state. We choose a length of time such that the fluctuations in number of patients in the system is minimized beyond the warm-up period. A sample of the charts used in choosing this warm-up period length are included in Appendix C. Additionally, we only consider patients who complete service before the end of the 365-day period, so as to remove bias from the cool-down of the system. Consequently, we calculate performance statistics using data from patients that enter the system after Day 15 and leave the system before Day 365.

3.5.2 Parameter Estimation

To test the performance of the algorithms under real world settings, we estimate parameter values using data provided by Children's where possible. Specifically, we use 2011 inpatient volume data, unit specifications, and documented overflow policies. These overflow policies identify the ideal or preferred unit by patient type, which overflow units a patient can be feasibly assigned to, and the preferences among these units. For each patient type we define primary, secondary, and tertiary units to which patients can be assigned.

A summary of the base case parameters is provided in Table 6. Values for I, J, and N_j are derived directly from Children's data. Due to a lack of available data the values N_0 and μ_i^0 are approximated. The limitations of these approximations are discussed in Section 3.6. Patient overflow protocol is used in the definition of $a_{i,j}$. Using Children's policies, we convert the overflow rules into numeric valuations. The weighting of the value achieved from assignment is in direct contrast with the daily cost of boarding.

Table 6: System parameters for base case

Parameter	Simulation Values
I	11
J	9
N_0	35
N_j	{14, 21, 18, 17, 28, 26, 27, 10}
μ_i^0	Assuming the average utilization of upstream unit by non-boarding patients is 83 %, all patient types assumed to have an average service time of .5 days.
μ_i	Six sets of downstream service rate parameters. See Tables 7 and 8 for the full definition.
$a_{i,j}$	Daily value of 5 if unit j is a primary unit for patient i. Values of 3 and 1 if unit j is secondary or tertiary unit, respectively, for patients of type i.
b_i	Daily value of 2 for all patient types.

The length of stay, and associated service rate μ_i are derived using information obtained from the overflow policies. We create six sets of parameters for μ_i, representing the patient service time, by varying both the relationship between supply and demand in the downstream units (SET1, SET2, SET3) and scaling parameters to

represent high and low utilization of the downstream units (HI and LO). A description of the relationship between the six parameter sets (S1LO, S1HI, S2LO, S2HI, S3LO, S3HI) is provided in Table 7.

Table 7: Description of six parameter sets (μ_i)

Match of Supply and Demand in Downstream Units	High Utilization (HI)	Low Utilization (LO)
SET1: Proportional in primary unit	S1HI	S1LO
SET2: Proportional in primary and secondary units	S2HI	S2LO
SET3: Proportional in all feasible units	S3HI	S3LO

Ideally the distribution of beds among hospital units is proportional to the flow of patients into their preferred units (SET1), but often the allocation of beds may be mismatched due to changes in patient mix over time. For a more robust comparison of the performance of the algorithms, we create two downstream service rate parameter sets that vary the match of patient flow and unit capacity. In the second set, SET2, for each patient class we divide patient flow evenly among the primary and secondary units and proceed in a similar fashion as above. Lastly, service rates for SET3 are calculated similarly dividing patient flow evenly among all feasible units.

This calculation of average service time by dividing the available capacity among the expected arrivals results in a system with very high utilization (HI), as it leaves little availability of excess capacity. To represent a system with lower utilization (LO), we create similar parameters in which the expected service time is 80% of that used in the high utilization parameter set. The average downstream length of stay for each patient type in these six instance sets is presented in Table 8.

We create 20 instances for each of the 6 parameters sets for use in our analysis and comparison of the algorithms performance.

Table 8: Average length of stay by patient class for parameter sets: S1LO, S2LO, S3L0, S1HI, S2HI, and S3HI

Patient Class	Average Length of Stay in Downstream Unit (days)					
	S1LO	S2LO	S3LO	S1HI	S2HI	S3HI
1	4.5	4.1	5.3	5.7	5.2	6.7
2	8.8	3.3	5.3	11	4.2	6.6
3	3.5	5	5.8	4.4	6.3	7.3
4	8.8	4.1	6	11	5.2	7.5
5	4.1	5	4.5	5.2	6.2	5.6
6	4.3	4.9	5	5.4	6.1	6.3
7	8.6	6.2	5.5	10.7	7.1	6.8
8	5.3	4.6	6.6	6.6	5.8	8.2
9	3.5	4.7	5.8	4.4	5.9	7.3
10	4.9	6.2	4.7	6.1	7.7	5.9
11	1.5	1.5	1.1	1.9	1.9	1.4

3.5.3 Optimality Gap Reduction

To illustrate the performance of the algorithms independent of the effect from patient diversions, we first examine conduct a numerical analysis of the optimality gap reduction in systems without patient diversion.

Without knowledge of the exact optimal solution, ensuring that the optimal solution has no diversions is difficult. If the capacity of the upstream unit is greater than the maximum number of patients that can be in the system at any time, as defined by the specific instance, it is guaranteed that no patients are diverted from the system under any policies.

We develop "augmented" instances in which the capacity of the upstream unit is increased to ensure no patient diversion. While this does not represent the system characteristics seen in practice, a calculation of the upper bound in these augmented instances, with expanded upstream unit capacity, allows for a better understanding of the performance improvement achieved through use of the assignment algorithms in comparison with the myopic approaches.

By augmenting the instances in sets S1LO, S1HI, S2LO, S2HI, S3LO, and S3HI we

compare the optimality gap for each algorithm with the optimality gap for the myopic approaches, GREEDY and VALONLY. We calculate the ratio of these optimality gaps individually for each instance. Table 9 includes the total reward and optimality gap reduction with respect to the GREEDY approach for each instance in S3HI. Because the UB algorithm produces an upper bound of the optimal reward, the optimality gap reduction listed here is an upper bound on the true optimality gap reduction.

Table 9: Optimality gap reduction in comparison to GREEDY for 20 instances in S3HI

	Total Reward					Optimality Gap Reduction		
INSTANCE	GREEDY	UB	PROB	STOCH	HYBRID	PROB	STOCH	HYBRID
0	196,016	235,651	215,110	213,259	213,690	48%	44%	45%
1	200,775	238,005	213,616	211,190	211,201	34%	28%	28%
2	197,636	236,058	217,204	213,661	213,916	51%	42%	42%
3	198,144	235,258	214,625	212,996	213,785	44%	40%	42%
4	197,795	236,774	212,967	209,428	209,989	39%	30%	31%
5	202,772	235,546	208,989	209,194	209,373	19%	20%	20%
6	206,003	237,479	214,749	214,254	214,379	28%	26%	27%
7	200,254	237,579	213,484	211,692	212,738	35%	31%	33%
8	199,974	237,591	212,251	212,493	211,936	33%	33%	32%
9	197,847	239,610	212,800	210,612	210,976	36%	31%	31%
10	195,158	239,738	217,949	215,182	214,817	51%	45%	44%
11	200,068	239,131	214,601	214,859	215,997	37%	38%	41%
12	198,106	236,261	211,911	211,417	211,467	36%	35%	35%
13	202,377	237,309	215,400	215,667	215,199	37%	38%	37%
14	201,056	237,570	211,851	212,168	212,766	30%	30%	32%
15	200,563	236,800	213,855	212,083	213,004	37%	32%	34%
16	200,518	239,114	215,537	212,692	212,279	39%	32%	30%
17	196,820	234,969	212,764	210,875	211,458	42%	37%	38%
18	203,275	238,467	216,222	215,774	216,402	37%	36%	37%
19	196,496	237,165	213,205	212,671	213,012	41%	40%	41%
Average						38%	34%	35%

As shown in Table 9, in comparison with the GREEDY method, a significant decrease in the calculated optimality gap is achieved in each instance in the set S3HI through use of any of the three algorithms: PROB, STOCH, HYBRID. For example, the HYBRID algorithm results in an average optimality gap reduction of 35%, with the individual optimality gap reductions ranging between 20% and 45% among all instances.

While all three algorithms achieve a higher total reward than GREEDY for all instances in S3HI, this finding does not hold for all instance sets and when comparing against the performance of the VALONLY approach. A summary of the mean,

minimum, and maximum optimality gap reductions with respect to the performance with GREEDY and VALONLY for all six instance sets is provided in Table 10.

Table 10: Mean, minimum, and maximum optimality gap reductions with respect to GREEDY and VALONLY

	Optimality Gap Reduction - GREEDY			Optimality Gap Reduction - VALONLY		
	PROB	STOCH	HYBRID	PROB	STOCH	HYBRID
S1LO	-16% [-41% , 1%]	31% [24% , 43%]	31% [25%, 41%]	-18% [-41% , 0%]	30% [21% , 42%]	30% [21% , 41%]
S1HI	41% [30% , 50%]	36% [22% , 45%]	37% [25% , 47%]	9% [2% , 16%]	1% [-6% , 7%]	3% [-4% , 10%]
S2LO	-4% [-19% , 7%]	29% [23% , 34%]	30% [22% , 36%]	-7% [-23% , 8%]	27% [17% , 34%]	27% [17% , 36%]
S2HI	43% [27% , 56%]	39% [24% , 53%]	39% [21% , 52%]	11% [6% , 19%]	5% [-4% , 14%]	5% [-3% , 12%]
S3LO	-3% [-9% , 4%]	21% [15% , 26%]	21% [16% , 26%]	-6% [-12% , 3%]	19% [15% , 24%]	19% [14% , 25%]
S3HI	38% [19% , 51%]	34% [20% , 45%]	35% [20% , 45%]	8% [0% , 21%]	2% [-7% , 9%]	4% [-5% , 13%]

For the instances with low utilization (S1LO, S2LO, and S3LO) PROB performs worse than both myopic approaches on average, but STOCH and HYBRID assignment algorithms result in significant improvements over both myopic approaches. Alternatively, for the high utilization instance sets, on average PROB results in the greatest optimality gap reduction, but all three algorithms improve the optimality gap over that of GREEDY for all 60 instances. Similarly, PROB performs better than VALONLY for all 60 instances.

We hypothesize that the poorer performance of PROB in low utilization instances is due to the greater value placed on waiting for a better assignment. By definition, as more downstream beds are occupied, the probability of a bed becoming available receives more weight and thus provides more incentive for boarding.

While these results are for the augmented instances with a large capacity in the upstream unit, similar results regarding the relative performance of the algorithms in high and low utilization systems occurs in the original instances. As a result, we separate our discussion of the performance of the algorithms below with respect to the level of utilization. In this analysis we explore the impact of the original upstream

capacity constraints on other system objectives.

3.5.4 Multi-Objective Analysis

From the computational results in Section 3.5.3 we gain an understanding for the performance of the algorithms for instances in which the upstream unit is not a limiting factor and no patients are diverted from the system. In the remainder of the computational study, we refer to the original instances in which the upstream unit is a limiting factor. In systems with capacitated upstream units bed managers must consider how assignment decisions impact the availability of beds in the upstream units and the resulting diversion rates.

Utilizing the discrete event simulation, we examine the performance of these algorithms with respect to the metrics defined in Table 5. Since we are only concerned with algorithms which outperform myopic approaches on average, we limit the tests to the STOCH and HYBRID algorithms for low utilization instances (S1LO, S2LO, and S3LO).

3.5.4.1 Performance in Low Utilization Instances

The results for STOCH and HYBRID compared with GREEDY and VALONLY are provided in Tables 11 and 12, respectively. In the low utilization systems, through the use of STOCH or HYBRID simultaneous improvements in all performance metrics is achieved on average. The improvement among all metrics with STOCH and HYBRID is a result of the consideration of the interactions between different patient types through proactive assignment to off-service units ensuring timely assignments for all patients. Utilizing a paired two-sample t-test ($\alpha = .05$) we confirm that the average performance of both STOCH and HYBRID is significantly different than the average performance of the same metric achieved with the myopic approaches. While the difference in the average median assignment times is statistically significant, due to small median assignment times in all instances the difference is not practically

significant.

Table 11: Average performance of STOCH and HYBRID relative to GREEDY among 60 low utilization instances (S1LO, S2LO, and S3LO)

		Total Reward	Assignment Value	Boarding Cost	Mean Assignment Time (mins)	Median Assignment Time (mins)
S1LO	GREEDY	207,650.7	208,112.4	461.5	32.8	1.19E-12
	STOCH	+1.26%	+1.24%	-7.45%	-7.47%	-11.90%
	HYBRID	+1.25%	+1.23%	-7.27%	-7.28%	-11.53%
S2LO	GREEDY	204,733.4	205,318.8	585.3	41.7	1.44E-12
	STOCH	+1.39%	+1.36%	-9.64%	-9.64%	-12.06%
	HYBRID	+1.41%	+1.38%	-9.44%	-9.44%	-11.89%
S3LO	GREEDY	203,500.9	203,847.9	347.0	24.7	9.62E-13
	STOCH	+1.09%	+1.06%	-11.88%	-11.89%	-13.26%
	HYBRID	+1.08%	+1.06%	-11.24%	-11.25%	-14.18%

Table 12: Average performance of STOCH and HYBRID relative to VALONLY among 60 low utilization instances (S1LO, S2LO, and S3LO)

		Total Reward	Assignment Value	Boarding Cost	Mean Assignment Time (mins)	Median Assignment Time (mins)
S1LO	VALONLY	207,795.2	208,267.4	472.3	33.6	1.19E-12
	STOCH	+1.19%	+1.16%	-9.56%	-9.58%	-12.37%
	HYBRID	+1.18%	+1.15%	-9.39%	-9.40%	-12.00%
S2LO	VALONLY	205,018.9	205,617.6	598.7	42.7	1.46E-12
	STOCH	+1.25%	+1.21%	-11.67%	-11.68%	-13.67%
	HYBRID	+1.27%	+1.23%	-11.47%	-11.49%	-13.50%
S3LO	VALONLY	203,761.8	204,132.6	370.8	26.4	9.97E-13
	STOCH	+0.96%	+0.92%	-17.53%	-17.54%	-16.32%
	HYBRID	+0.95%	+0.91%	-16.94%	-16.95%	-17.21%

Additionally, these algorithms impact the rate at which patients are assigned to on-service units and the rate at which patients do not receive assignment and exit directly from the upstream unit. The distribution of the percentage of patients in each category for low utilization instances is provided in Table 13. Use of HYBRID or STOCH results in improved performance with a higher rate of on-service assignment and lower rates of off-service assignment, upstream departures, and diversions.

Due to the low utilization in the instances for sets S1LO, S2LO, and S3LO, as seen in the results of the simulation even with a capacitated upstream unit the average diversion rate is approximately zero for all instance sets and the algorithms have no significant impact on diversions. This implies that for the low utilization instance sets the optimality gaps for the augmented instances (Table 10) provide a reasonable approximation of the true optimality gap reduction in the original instances.

Table 13: Distribution of patients among on-service, off-service, upstream departures and diversions in 60 low utilization instances (S1LO, S2LO, and S3LO)

		On-Service Assignments	Off-Service Assignments	Upstream Departures	Diversions
S1LO	GREEDY	90.91%	7.946%	1.147%	0.000%
	VALONLY	91.05%	7.775%	1.178%	0.000%
	STOCH	93.91%	4.973%	1.115%	0.000%
	HYBRID	93.90%	4.980%	1.118%	0.000%
S2LO	GREEDY	88.09%	10.373%	1.538%	0.002%
	VALONLY	88.41%	10.038%	1.545%	0.005%
	STOCH	90.80%	7.756%	1.441%	0.003%
	HYBRID	90.81%	7.734%	1.452%	0.003%
S3LO	GREEDY	88.10%	11.058%	0.841%	0.003%
	VALONLY	88.46%	10.636%	0.907%	0.002%
	STOCH	90.40%	8.783%	0.816%	0.002%
	HYBRID	90.38%	8.815%	0.805%	0.002%

Observation 3.5.1. *In low utilization systems, simultaneous improvement in all performance metrics and a decrease in the optimality gap can be accomplished through the implementation of improved bed assignment practices.*

3.5.4.2 Performance in High Utilization Instances

As demonstrated with the upper bound analysis (Table 9), all three algorithms perform better than the myopic approaches in the high utilization instances with respect to the average total reward achieved, but the performance of the algorithms with respect to the other metrics varies. None of the algorithms experience an improvement in all performance metrics across all three instance sets.

The average performance of the algorithms in the high utilization instance sets (S1HI, S2HI, S3HI) are presented in Tables 14 and 15. Again, a two-sample paired t-test confirms the significant difference of the average performance under each algorithm and each of the myopic approaches. The table is annotated to display when the difference is not significant.

PROB performs better with respect to the average total reward and assignment

value and worse with respect to the assignment times in comparison to STOCH and HYBRID. While the STOCH algorithm achieves improvements in assignment times due to the proactive off-service assignment of patients, the PROB algorithm achieves higher total assignment value through proactive boarding, in which a patient waits for a better bed or beds are "reserved" for future arrivals. This practice of proactive boarding results in increased assignment times.

STOCH and HYBRID perform similarly due to the similar structures, although HYBRID consistently outperforms STOCH with respect to total reward. HYBRID consistently results in the lowest average diversion rates among the algorithms, having no significant difference from the myopic approaches. For high utilization instances, each algorithm achieves the best average performance for one of the metrics under consideration, but none have the best performance among all metrics.

Observation 3.5.2. *For the high utilization systems, simultaneous improvement in all performance metrics is not possible with the algorithms we developed. Instead, a bed manager must choose the bed assignment approach that achieves an appropriate tradeoff among the different objectives.*

The rates of on-service and off-service assignment for each algorithm are provided in Table 16. PROB, STOCH, and HYBRID all perform similarly in their distribution of patients although they perform differently in total reward and mean and median assignment times. With the STOCH algorithm performing better than PROB in assignment times and the PROB performing better than STOCH in assignment value, the performance of the HYBRID algorithm is a compromise between the two for all instance sets and metrics, since the increase in median assignment time has not been shown to be statistically significant.

Observation 3.5.3. *HYBRID balances the high total assignment value from the PROB with the low boarding costs achieved by STOCH. Regardless of the utilization*

level, HYBRID performs well with respect to both the primary objectives of maximizing value from appropriateness of assignments and minimizing the mean assignment time, as well as secondary objectives such as number accepted and median assignment time.

As a result of the intrinsic tradeoffs in the metrics, the choice of algorithm depends on the preferences by the hospital staff.

Table 14: Average performance of algorithms relative to GREEDY among 60 high utilization instances (S1HI, S2HI, S3HI) (*: denotes no significant difference between GREEDY and algorithm)

		Total Reward	Assignment Value	Boarding Cost	Mean Assignment Time (mins)	Median Assignment Time (mins)	Accepted
S1HI	GREEDY	203,617.75	207,615.20	3,997.50	289.05	7.36	9,952.80
	PROB	+7.11%	+6.93%	-2.16%	-2.10%	+31.40%	-0.06%
	STOCH	+6.46%	+6.28%	-2.55%	-2.55%	+6.21*%	0.00*%
	HYBRID	+6.68%	+6.50%	-2.35%	-2.36%	+10.13*%	+0.01*%
S2HI	GREEDY	205,245.40	209,318.15	4,072.59	294.68	8.30	9,945.15
	PROB	+5.82%	+5.63%	-3.78%	-3.76%	+42.32%	-0.03%
	STOCH	+5.23%	+5.07%	-2.87%	-2.87%	-2.11*%	0.00*%
	HYBRID	+5.33%	+5.17%	-2.79%	-2.79%	+10.77*%	-0.01*%
S3HI	GREEDY	201,086.70	204,774.65	3,687.82	266.99	13.78	9,939.70
	PROB	+6.11%	+5.97%	-1.70%	-1.66%	+39.87%	-0.04%
	STOCH	+5.40%	+5.27%	-2.06%	-2.05%	+11.91*%	-0.01*%
	HYBRID	+5.52%	+5.38%	-2.52%	-2.51%	+18.85*%	0.00*%

Table 15: Average performance of algorithms relative to VALONLY among 60 high utilization instances (S1HI, S2HI, S3HI) (*: denotes no significant difference between VALONLY and algorithm)

		Total Reward	Assignment Value	Boarding Cost	Mean Assignment Time (mins)	Median Assignment Time (mins)	Accepted
S1HI	VALONLY	215,508.60	219,496.10	3,987.27	288.36	9.47	9,951.25
	PROB	+1.20%	+1.14%	-1.90%	-1.86%	+2.08*%	-0.04%
	STOCH	+0.58%	+0.53%	-2.30%	-2.31%	-17.49%	+0.01%
	HYBRID	+0.79%	+0.74%	-2.10%	-2.12%	-14.44%	+0.02%
S2HI	VALONLY	214,827.30	218,826.40	3,998.97	289.36	8.81	9,944.90
	PROB	+1.10%	+1.04%	-2.01%	-1.98%	+34.02%	-0.03%
	STOCH	+0.54%	+0.51%	-1.08%	-1.08%	-7.81*%	0.00*%
	HYBRID	+0.63%	+0.60%	-1.00%	-1.00%	+4.31*%	0.00*%
S3HI	VALONLY	211,334.85	214,993.25	3,658.52	264.90	14.53	9,938.95
	PROB	+0.97%	+0.94%	-0.91%	-0.88%	+32.64%	-0.03%
	STOCH	+0.29%	+0.26%	-1.28%	-1.28%	+6.12*%	0.00*%
	HYBRID	+0.41%	+0.37%	-1.74%	-1.75%	+12.70*%	+0.01*%

3.5.5 Sensitivity to Cost Parameters

To examine the impact of the cost of boarding, b_i, on algorithm performance we focus our study on the instances most likely to represent a hospital system, with

Table 16: Distribution of patients among on-service, off-service, upstream departures and diversions in 60 high utilization instances (S1HI, S2HI, S3HI)

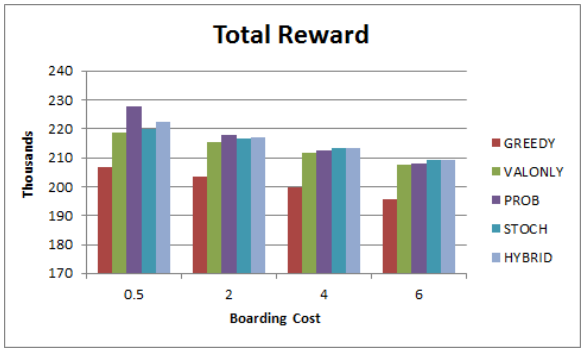
		On-Service Assignments	Off-Service Assignments	Upstream Departures	Diversions
S1HI	GREEDY	66.31%	26.61%	6.06%	1.02%
	VALONLY	72.29%	21.00%	5.67%	1.03%
	PROB	73.71%	19.67%	5.53%	1.08%
	STOCH	73.40%	19.97%	5.61%	1.02%
	HYBRID	73.52%	19.83%	5.63%	1.02%
S2HI	GREEDY	63.32%	29.00%	6.59%	1.08%
	VALONLY	68.57%	24.34%	6.01%	1.08%
	PROB	69.44%	23.67%	5.78%	1.11%
	STOCH	69.33%	23.64%	5.95%	1.08%
	HYBRID	69.42%	23.54%	5.96%	1.08%
S3HI	GREEDY	61.81%	30.92%	6.37%	0.90%
	VALONLY	67.65%	25.71%	5.74%	0.91%
	PROB	68.77%	24.71%	5.58%	0.94%
	STOCH	68.18%	25.25%	5.66%	0.91%
	HYBRID	68.35%	25.13%	5.62%	0.90%

well matched capacity and demand in the downstream units (S1LO and S1HI). For each instance we vary the daily boarding cost to be .5, 2, 4, and 6. The assignment values remain the same with a daily value of 5 received from assignment to the primary unit and daily values 3 and 1 for assignment to secondary and tertiary units, respectively. In Figure 14 we show how the change in b_i impacts the average total reward, assignment value, and mean and median assignment times for instances in set S1HI. Similar results occur for instance set S1LO.

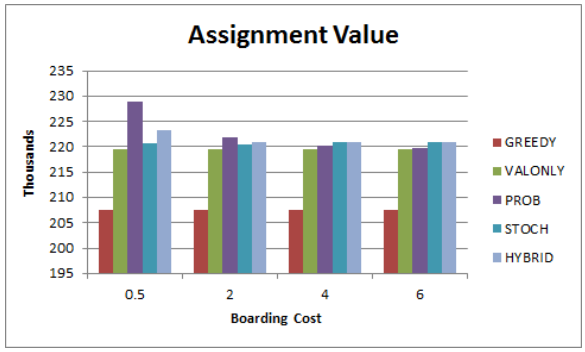
For GREEDY, VALONLY, STOCH, and HYBRID the relative performance with respect to total reward, assignment value, mean assignment time, and median assignment time do not significantly change with increasing boarding costs. With PROB, while the average median assignment time is not affected, the average assignment values and mean assignment times decrease with increases in boarding costs. These changes cause the total reward to decrease in a similar fashion.

PROB outperforms the other algorithms, including the myopic ones, for low boarding costs, but the performance relative to the other algorithms is poorer with increases

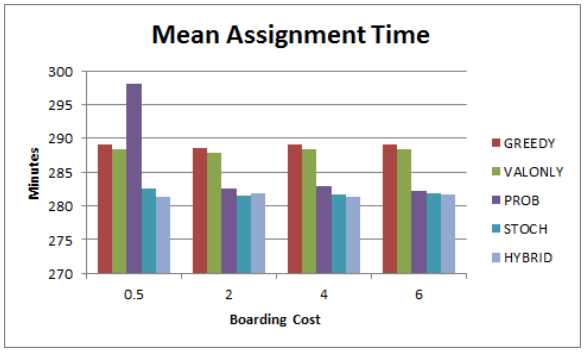
Figure 14: Effect of changes to b_i on average total reward, assignment value, mean assignment time, and median assignment time for instance set S1HI



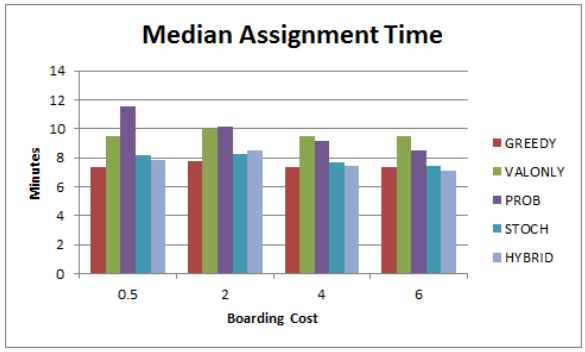
(a) Average Total Reward



(b) Average Assignment Value



(c) Average Mean Assignment Time



(d) Average Median Assignment Time

in the boarding cost. We hypothesize that because PROB places more value on nonassignment than the other algorithms, when boarding costs are low and postponement of assignment is naturally favored PROB outperforms other algorithms.

The average total reward in all of the algorithms converge to a similar level as the boarding cost increases. If boarding costs are high, the objective is to assign patients as quickly as possible regardless of the appropriateness of the bed. Since the algorithms we develop are structured to focus on the appropriateness of the assignments, it follows that they perform better than the myopic approaches in systems with smaller boarding costs relative to assignment values. Similarly the optimality gaps of the algorithms in the augmented instances converge as b_i increases.

Observation 3.5.4. *If a hospital has a significant preference for decreased assignment times over increased on-service assignments VALONLY performs similarly to*

the other algorithms. As a result, the value of new bed assignment practices is minimal is such a setting.

3.5.6 Impact of Weighting of Patient Preferences

To evaluate the sensitivity of algorithm performance relative to the assignment values, we reassign the daily value of assignment to a preferred unit to be 9, rather than 5. We choose 9, such that the values increase exponentially, rather than linearly. Thus we examine how an increased preference for patients to be assigned to primary units impacts the algorithm performance. We evaluate this performance with hourly boarding costs of .5, 2, 4, and 6.

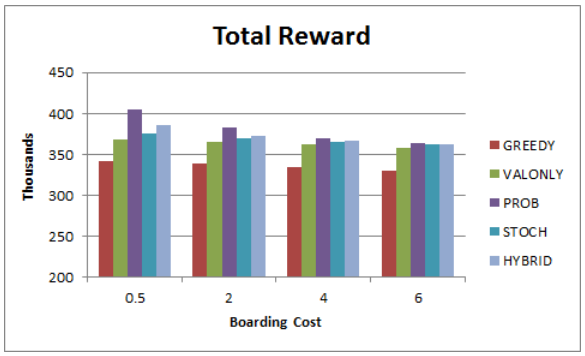
The impact of this change is demonstrated through comparison of Figures 14 and 15. The average mean assignment times change only slightly in response to an increase in the preferred unit assignment value. The average mean and median assignment times decrease when the boarding cost is .5 and increases for all higher boarding cost levels. The assignment values and total reward for the algorithms increase proportionally with the change to the assignment values.

Observation 3.5.5. *The performance of the HYBRID algorithm with respect to other algorithms, specifically PROB and STOCH, is robust to changes in boarding costs and assignment value, although the relative performance with respect to the myopic approaches decreases as boarding costs increase.*

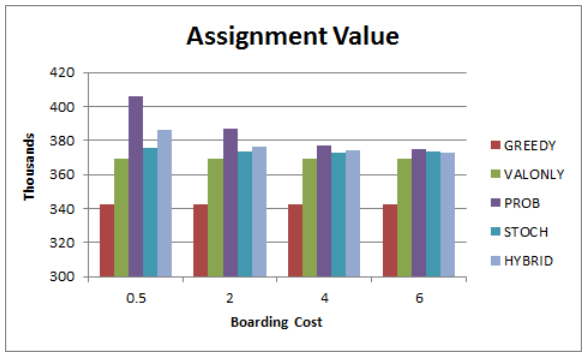
3.6 Future Work

While we have demonstrated the value of implementing analytical methods for bed management, opportunities exist to improve the models and algorithms presented above. Additional work is needed to identify structural properties of the MDP presented in Section 3.3, which can in turn be used to improve the current bed assignment algorithms and to develop new bed assignment algorithms, possibly through the use

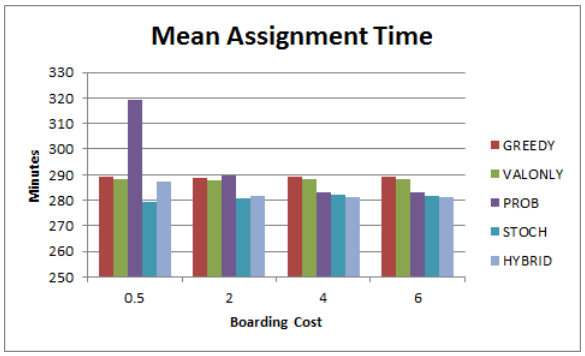
Figure 15: Effect of boarding costs on average total reward, assignment value, mean assignment time, and median assignment time for instance set S1HI with increased value of preferred assignments



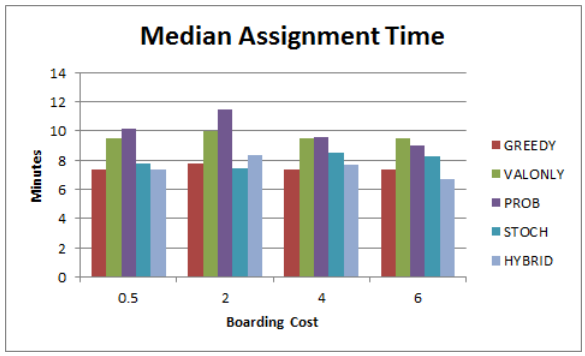
(a) Average Total Reward (New $a_{i,j}$ Values)



(b) Average Assignment Value (New $a_{i,j}$ Values)



(c) Average Mean Assignment Time (New $a_{i,j}$ Values)



(d) Average Median Assignment Time (New $a_{i,j}$ Values)

of approximate dynamic programming. These structural properties may also prove useful in examining the best allocation of beds among units.

Due to the limitations of the approximations for average service times in the upstream and downstream units, additional data collection is needed to further improve the computational study. Additionally, testing of the bed assignment algorithms in a variety of hospital settings will allow for further refinement of the bed assignment algorithms.

In addition to improvements to the current algorithms, further research is needed to extend our models and methods. First, an extension that incorporates multiple upstream units can be examined. Assuming that all of the upstream units are large enough such that no patients are diverted from the hospital, the models developed

above provide an accurate representation of the true system and result in similar performance with respect to both primary and secondary metrics. When there are multiple upstream units for which the capacities are constrained resulting in diverted patients, a single upstream unit, which assumes pooled resources, will not accurately represent the diversion rates. To address these capacity constraints, the algorithms need to directly incorporate the state of available beds. The algorithms developed above are not designed with the direct intention of reducing patient diversions, although we have explored the impact with respect to this objective.We leave the consideration of the interactions between admission control policies and bed assignment policies for future research.

We hypothesize that one of two extensions can be developed to address the effects of constrained capacity in multiple upstream units. First, the addition of a policy for choosing among similar patients in different upstream units based on the state of the system could be considered. Alternatively, we hypothesize that the algorithms and models presented above can be adapted to include state-dependent boarding costs, rather than constant costs. These state-dependent boarding costs would inform preferences for patients in different units, accounting for the instantaneous utilization of the units. We leave these extensions and the development of appropriate policies or state-dependent boarding cost functions for future research.

Second, while the assumptions of Poisson arrivals and exponential service times have been shown to be reasonable in hospital settings, extensions that relax these assumptions and test the performance of the algorithms under alternate assumptions for arrival and departure patterns would be valuable. This may include changes to the structure of the expectations for arrivals and departures used as input to the algorithms. For example, instead of knowing the expected length of stay for all patients at the time of arrival, often the expectations about the patient departures change over time. Thus, we can examine the performance of algorithms considering

changing expectations about departure and assignment requests. The structure of the STOCH, PROB, and HYBRID algorithms, while developed with assumptions of Markovian arrivals and service times, do not explicitly require these assumptions for implementation and their performance should be examined in systems with alternate assumptions.

Both of these extensions have potential to further the readiness for application of dynamic bed assignment algorithms and implementation of decision support tools in hospitals for use by bed management staff. Finally, additional work is needed to test these algorithms with data collected from other hospital systems to examine the robustness with respect to model parameters and hospital policies.

3.7 Conclusion

In this chapter we developed a queueing framework and MDP model for hospital bed management addressing the real-time dynamic assignment of patients to hospital beds. Utilizing the features of the queueing framework, we developed three algorithms for aiding in bed assignment decisions using real-time information provided by bed management information systems. Each of these algorithms incorporate the stochastic features of the hospital processes in the process of choosing assignments. With a simulation, we tested the performance of these algorithms against standard practices. Additionally, we developed an algorithm to calculate an upper bound on system performance to further evaluate the value of the algorithms.

Through comparison with myopic algorithms, we identified the opportunity and ability to improve bed assignment practices and support bed management teams in hospitals. We improve these practices through the use of real time information about bed requests and utilization found in bed management information systems. With the development of the PROB and STOCH algorithms, we demonstrated the application of new algorithms that utilize the probabilistic and stochastic system characteristics

to aid in state-dependent routing. With the HYBRID algorithm we developed a policy that harnessed the strengths of two algorithms and stochastic optimization techniques. While PROB performed well with respect to the appropriateness of the bed assignment at the expense of timeliness and STOCH performed well with respect to timeliness, the HYBRID algorithm achieved a balanced improvement with respect to both appropriateness and timeliness and closed the optimality gap.

With the HYBRID algorithm we provide an example of how two fields within stochastic optimization, including stochastic optimization models with recourse and MDP derived transition probabilities, can be combined to aid in control of a complex queueing system. This is the first work to our knowledge to address the development of state-dependent routing policies in a queueing network with multiple customer classes, cross-trained servers, and in which service rates are dependent on customer class rather than the server pool. Additionally, our focus on the appropriateness of assignments and the assumption that the time in system is not influenced by the routing policies further distinguishes our problem of study. While future work is needed to extend these models for implementation, we make contributions by defining and modeling the dynamic assignment problem in the hospital bed management setting and developing analytical methods to address the operational component of these processes. We showed the value of continued research in this problem setting.

CHAPTER IV

PATIENT REDIRECTION IN HOSPITAL SYSTEMS WITH PREFERENCE CONSTRAINTS

4.1 Introduction

As discussed in Chapter 3, many hospitals are often plagued with overcrowding and bed shortages in both the emergency department (ED), as well as the inpatient units [13, 21, 33, 67]. To manage and alleviate overcrowding, hospital managers may implement policies that address both the rate of external arrivals as well as internal bed assignment operations. In the previous chapter, we developed models and methods to address the internal hospital policies of assigning and allocating beds to admitted patients. In this chapter we address the problem of managing patient arrivals to internal hospital units (non-ED), through redirection of patients to other hospitals.

We focus our attention on the development of policies for redirection of patients to other hospitals, e.g., in the same system, in the face of overcrowding or excess demand at one location. The ability to control flow of non-emergent patients into the hospital system is most applicable when the hospital is a member of a system of hospitals. In this case, patients seeking to be admitted to a hospital unit that is overloaded may be safely redirected to another hospital in the same system. In a hospital that functions independently, and does not have access to excess capacity, this form of admission control may not be feasible.

We address the definition of redirection policies for one hospital unit, in which all beds are equivalent, and from which multiple patient types request care. Upon request, based on the unit census, or the number of occupied beds, and the type of patient making the request, a decision is made to admit the patient to the hospital

unit or redirect to a similar unit in another hospital. Given the structure of the policies currently used in hospitals, we assume that the hospital utilizes a threshold policy, in which each patient type has a corresponding threshold such that once the total number of occupied beds in the unit reaches the threshold, patients of that type are redirected. This structure ensures equity in the treatment of similar patients who arrive to the system in a similar setting. We assume there is no queue for patients to enter the unit. Therefore, if the unit is at capacity, all patients are redirected, or diverted, regardless of patient type.

While we assume that units in different hospitals are similar (e.g., a focus on cardiology or neurology), we address the fact that some hospitals may have specialized equipment or facilities that are necessary for treating a subset of the total patient population. Therefore, for some patients only specific hospitals may be appropriate. As such, we do not treat the redirection of all patient types as equivalent and we assume that each patient type has a corresponding penalty representing the costs of redirection to another hospital.

Our focus on developing patient type-specific threshold redirection policies is driven by discussions with administrators in a hospital system which seeks to balance patient flow within the hospitals and to ensure that patients receiving specialized care are not turned away due to lack of capacity. In alignment with the mission and goals of the hospital, we assume the relative preferences for redirection and the order with which patients are eligible for redirection is fixed, and we therefore seek to identify the appropriate redirection thresholds given these constraints. In other words, the hospital policies state that patients of a particular type can not be redirected unless patients of a lower priority are also being redirected, thereby defining preferences for redirection of patients. We model these redirection preferences with corresponding penalties and redirection costs.

While ensuring that patients receive appropriate care is the first priority, there is

an objective to minimize the number of patients requiring specialized services that are redirected. Additionally, the administrators seek to efficiently utilize hospital resources. Due to the multiple interests of the hospital, we examine three objectives: (i) to ensure that beds are available for patients requiring care at a specific facility through minimizing the total redirection costs incurred for all patient types, (ii) to maximize the use of hospital beds, and (iii) to minimize the likelihood of a unit being fully occupied, in which case complete diversion of all patients is necessary. To identify threshold policies that simultaneously address these three objectives while satisfying the preference constraints, we model the system as a Markov chain defined by the redirection threshold policy.

In the remainder of this chapter we first present a review of relevant literature in Section 4.2. Next we describe our model of the system with patient redirection policies and preference constraints in Section 4.3. We discuss the impact of incremental changes to the threshold policies and identify necessary conditions for optimal patient redirection thresholds in Section 4.4. Finally, in Section 4.5 with numeric examples we illustrate the nature of tradeoffs in the three objectives and examine the influence of preferences on system performance, as defined by the objectives. In Section 4.6 we summarize the results from the analysis and discuss areas for future research.

4.2 Literature Review

The problem we address and the model we develop share similarities with previous work related to ambulance diversion (Section 4.2.1) and admission control policies in hospitals and other systems (Section 4.2.2). While models for ambulance diversion policies often differ in some of their objectives, the concept of identifying threshold policies in hospitals is directly relevant. Our focus on the appropriateness of the hospital unit by patient type and the costs associated with redirection is similar to models developed for the optimal control of queueing or loss systems.

4.2.1 Ambulance Diversion Models

The purpose of ambulance diversion policies is to address the overcrowding of emergency departments (EDs) by allowing for ambulances arriving to overcrowded EDs to be diverted to other local hospitals. Ideally the use of ambulance diversion policies can result in more balanced demand at EDs in a region and reduced patient wait times. Traditionally, ambulance diversion policies define a threshold such that when the occupancy of the ED exceeds this threshold, all incoming ambulances are diverted to other hospitals in the region, while other types of patients, such as walk-in patients, continue to be accepted until occupancy decreases to a specific level. With regard to ambulance diversion, most hospitals are classified as either "on diversion" or "off diversion", distinguishing those times when all ambulances are diverted or no ambulances are diverted. As a result, ambulance diversion policies identify criteria, or thresholds, for beginning diversion as well as ending diversion.

The impact of ambulance diversion policies has been well studied through case studies of hospital systems [3, 61, 64]. Multiple researchers have developed models to inform decision making for and to examine the impact of ambulance diversion policies. Using a simulation approach, Ramirez-Nafarrate, Fowler, and Wu (2010) demonstrate the impact of the choice of ambulance diversion policies for one hospital, including threshold triggers, on the performance metrics of time on diversion and average wait time for patients to enter the ED [71]. Expanding on this work, they use a simulation-optimization approach to examine a system of hospitals in which centralized decisions are made regarding diversion policies and destination policies for diverted ambulances, with the objective of minimizing non-value added time for patients such as the time in transport or waiting for admission [70].

Hagtvedt (2008) examines a similar problem of selective diversion of patients through a partial diversion strategy. In this problem a threshold is identified to determine when a portion of patients are diverted and assumes that the system goes

on "full diversion" once there is no additional capacity. A threshold is also identified for when to end full diversion [31]. With numeric experiments, he demonstrates how policies of partial ambulance diversion using threshold policies can minimize the time on full diversion and revenue loss while simultaneously benefiting hospitals, government, and the public. While we utilize a similar threshold model, assuming a birth-death process and no queue forming, we focus on a multiple threshold approach with patient specific costs dependent on the length of time being redirected. With our assumption of a system of hospitals, we do not consider payer type in the definition of redirection policies.

Expanding on the study of diversion policies in one hospital, Hagtvedt et al. (2009) utilize an agent-based model to examine partial ambulance diversion as a sequential game, with the purpose of identifying cooperative strategies. They note that the existence of hospital systems in which the interests of players are naturally aligned eases the ability to receive benefit from the implementation of diversion policies [32]. This reinforces our examination of redirection policies in a network of hospitals, where the incentives to improve patient care and overall system performance are aligned.

Deo and Gurvich (2011) examine the nature of ambulance diversion in a competitive environment through the use of a stylized model of ambulance diversion within a network of hospitals. They utilize a game-theoretic queueing framework to examine both centralized and decentralized settings with the goal of minimizing expected waiting time of patients [19]. Deo and Gurvich (2011) assume decision makers utilize threshold policies based on local hospital census information and that an infinite queue can form at EDs without the use of diversion policies [19].. With our focus on internal hospital units, we do not address queues forming for admittance.

Similar to the ambulance diversion models, we seek to utilize threshold policies to minimize the total time patients are diverted, or redirected, although we consider a more complex setting with multiple patient types with different costs of redirection.

Due to our focus on inpatient redirection, the objective of our model is both to reduce overcrowding and average cost of redirection, rather than reducing wait times which is more appropriate in an emergency setting.

4.2.2 Admission Control Models

In addition to models developed specifically for ambulance diversion, many researchers have examined problems of admission control, in which upon arrival of a customer a decision is made to admit or reject the customer based on the customer's characteristics and the state of the system. The patient redirection problem, with the focus on decision making based on patient type, is similar to admission control problems, particularly those in loss systems. A loss system is one in which customers are rejected from the system if there is no available capacity and does not allow a queue to form for service.

A summary of past models of admission control assuming multiple customer types within loss-systems is presented below. Many of these admission control problems occur in hospital settings, but do not specifically deal with ambulance diversion (Section 4.2.2.1). Other admission control models are developed for generic settings with multiple customer types (Section 4.2.2.2). The objectives of admission control vary with the setting.

4.2.2.1 Admission Control in Hospitals

Admission control in hospitals generally focuses on whether to accept a patient into a hospital based on the type of patient. These models assume patients arrive to the hospital and request service and a decision regarding admission is made in response [4, 29]. For example, Esogbue and Singh (1976) develop an admission control model with two patient classes and the goal of identifying a threshold after which only priority patients receive care. They seek to balance the penalties for turning away patients as well as the holding costs per unit time [20]. Similar to our model, they

assume that the cost of not fulfilling a patient request is dependent on the patient type. Additionally, they utilize a set of birth-death equations to represent the system and assume that the relative preference or priorities for rejecting patients is known. Dissimilarly, we examine a system with multiple patient types and a different set of objectives.

Helm et al. (2011) examine a system with three patient types using a Markov decision process model[36]. They develop a double threshold policy such that two of the patient types are denied hospital admission dependent on the system occupancy. The preferred patient type is always admitted as long as capacity is available. With the objective to balance costs of congestion and under utilization, they assume costs are independent of patient type [36]. In our model we also consider multiple threshold policies, with no limit on the number of thresholds. While our objectives of minimizing full diversion and maximizing efficiency are similar, we examine the tradeoffs in these objectives rather than assigning a cost for each objective. Additionally, we consider the relative importance of the different patient types, and do not consider the costs of diversions to be equivalent for all patient types.

4.2.2.2 Admission Control in Loss Systems

Örmeci, Burnetas, and van der Wal (2001) develop a stochastic model of admission control policies for a loss system with two customer classes, to address a wide variety of service and manufacturing systems [59]. They show the optimality of a threshold policy dependent on the number of customers in service. Rather than a penalty for rejecting a customer, they assume a reward is received upon completion of service. They maximize the total long-run average revenue over an infinite horizon. They show that the definition of a "preferred" customer class in the optimal solution is only guaranteed when specific system characteristics are met [59].

While Örmeci et al. (2001) allow for different service rates for different patient

types, the setting with equal service rates for all customers is studied by Miller (1969). Miller does not assume preferences among customer types and only examines the system with respect to the objective of maximizing revenue [56].

We seek to identify threshold policies for a system with more than two patient types and in which there is a preference for patients. With our assumption of different costs for each patient type, the objectives are aligned with admission control applications. Dealing with a hospital setting, and similar to the ambulance diversion problems, cost is a function of the length of time for which patients of a particular type are diverted, which is different than what is traditionally seen in admission control models. We build on existing literature by examining tradeoffs in multiple performance metrics in a loss system with more than two customer classes, with constraints on preferences for redirection and for which costs are dependent on patient type.

4.3 Model

We examine the development of admission control policies for a hospital unit with N beds and I patient types. Patients of type i arrive to the unit according to a Poisson process with rate λ_i. All patients, regardless of type, have an exponentially distributed service time with rate μ. Any patients arriving to the system when all beds are occupied are diverted from the system. Let $S = \{0 \dots N\}$ represent the set of all feasible states the unit may be in, as defined by the number of beds occupied, or the unit census. The long-run probability of being in each state is a function of the admission control policy, or patient redirection policy, in use.

We assume that preferences for patient redirection are fixed, such that a patient of type i can only be redirected if patients of type $i-1$ are also redirected, defining the preference for patient type i over patient type $i-1$. These preference constraints are aligned with the penalties or costs associated with redirection. A penalty or cost

k_i is incurred for each unit of time that patients of type i are redirected due to the admission control policy or lack of capacity. This measure of cost is in alignment with that seen in the ambulance diversion literature, in which the objective is to minimize the total time on diversion [31].

We consider admission control policies with a threshold structure, such that if the total number of occupied beds is greater or equal to a specific threshold, patients are redirected to another hospital. A redirection policy, or admission control policy, can therefore be defined with the vector $\vec{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_I)$. When the unit has Θ_i or more beds occupied, all patients of type i and patients with "lower preferences" are redirected upon arrival.

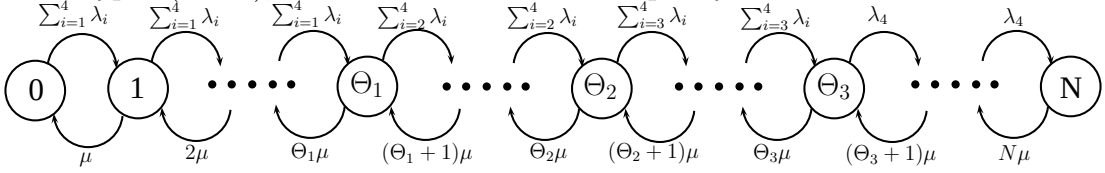
We seek to minimize the weighted sum of the penalties for redirection for the multiple patient types. In alignment with the preferences and the assumption that at least one patient is always accepted as long as a bed is available, we assume $k_1 \leq k_2 \leq \cdots \leq k_I$ and $\Theta_1 \leq \Theta_2 \leq \Theta_3 \leq \cdots \leq \Theta_I = N$.

Under a specific control policy the system can be modeled as a Markov chain. A depiction of the Markov chain, including states, transitions, and threshold policies is provided in Figure 16.

4.3.1 Limiting Probability Distributions

Under any feasible admission control policy, with the assumption that patients of type I are accepted as long as beds are available, there is a positive probability of visiting all states regardless of the initial state and therefore the Markov chain is recurrent

Figure 16: Description of Markov chain for patient redirection in a unit with N beds, 4 patient types ($I = 4$), and the admission control policy $\vec{\Theta}$



and irreducible, and as a result there is a unique limiting probability distribution [72, 73]. Let $P_n(\vec{\Theta})$ be the limiting probability of being in state n ($n \in S$) in the long run, or having n beds occupied when the admission control policy is $\vec{\Theta}$. We can characterize these limiting probabilities utilizing birth-death balance equations.

In a standard fashion, we define each $P_n(\vec{\Theta})$ by its relationship to $P_0(\vec{\Theta})$ and normalize the values by defining $P_0(\vec{\Theta})$ such that $\sum_{i=0}^{N} P_n(\vec{\Theta}) = 1$ [73]. For ease of notation in the definition of the limiting probabilities, we define ρ_j which is the ratio of the total arrival rate of all patients to the service rate of one patient in a state with Θ_j or more beds occupied (Equation 35).

$$\rho_j = \frac{\sum_{i=j}^{I} \lambda_i}{\mu} \qquad \forall j \in \{1 \dots I\} \tag{35}$$

Lemma 4.3.1. *Equations for evaluating the limiting probabilities for the $N+1$ states are provided in Equations 36 and 37.*

$$P_n(\vec{\Theta}) = \begin{cases} \frac{P_0(\vec{\Theta})\,{\rho_1}^n}{n!} & \forall n \in S : 0 < n \leq \Theta_1 \\ \frac{P_0(\vec{\Theta})\,{\rho_1}^{\Theta_1}\,{\rho_2}^{n-\Theta_1}}{n!} & \forall n \in S : \Theta_1 < n \leq \Theta_2 \\ \frac{P_0(\vec{\Theta})\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1}\,{\rho_3}^{n-\Theta_2}}{n!} & \forall n \in S : \Theta_2 < n \leq \Theta_3 \\ \dots & \dots \\ \frac{P_0(\vec{\Theta})\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1}\,{\rho_3}^{\Theta_3-\Theta_2}\dots{\rho_{I-1}}^{n-\Theta_{I-1}}}{n!} & \forall n \in S : \Theta_{I-1} < n \leq N \end{cases} \tag{36}$$

$$P_0(\vec{\Theta}) = \frac{1}{1 + \sum_{n=1}^{\Theta_1} \frac{\rho_1^n}{n!} + \sum_{n=\Theta_1+1}^{\Theta_2} \frac{{\rho_1}^{\Theta_1}\,{\rho_2}^{n-\Theta_1}}{n!} + \dots + \sum_{n=\Theta_{I-1}+1}^{N} \frac{{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1}\dots{\rho_{I-1}}^{n-\Theta_{I-1}}}{n!}} \tag{37}$$

Proof. The derivation of the equations for the limiting probabilities follow from the definition of the limiting probabilities in a birth-death process provided by Ross (1996)

in Equation (5.5.4) [72]. According to this equation, if the birth rate for some state n is b_n and the death rate is d_n, then the limiting probability of state n (P_n) is defined by the equation [72]

$$P_n = \frac{b_0 b_1 \dots b_{n-1}}{d_1 d_2 \dots d_n \left(1 + \sum_{n=1}^{\infty} \frac{b_0 b_1 \dots b_{n-1}}{d_1 d_2 \dots d_n}\right)}$$

We define the birth and death rates for each of the N+1 states in our system. For all states n such that $\Theta_{j-1} < n \leq \Theta_j$, the corresponding birth rate (b_n) is $\sum_{i=j}^{I} \lambda_i$. For each state n, the death rate (d_n) is $n\mu$. Utilizing these definitions of the birth and death rates and Equation (5.5.4) from Ross (1996) [72], Equations 36 and 37 follow intuitively with the assumption that $\sum_{i=0}^{N} P_n(\vec{\Theta}) = 1$. □

4.3.2 Performance Metrics

Utilizing the limiting probabilities of the Markov chain induced by the control policy, we define three performance metrics: average cost (per unit time), average utilization, and diversion rate.

Definition Average cost (per unit time), $K(\vec{\Theta})$, is the weighted sum of the penalties incurred from redirection of patients as a result of the patient redirection policy $\vec{\Theta}$ (see Equation 38). The definition of this metric accounts for the fact that the cost k_j is incurred for each unit of time spent in states n in which $n \geq \Theta_j$.

The mathematical definition of $K(\vec{\Theta})$ is provided in Equation 38, representing that the cost k_j is incurred for each unit of time that the system is in state $n \geq \Theta_j$. The average percentage of time spent in these states, $\sum_{n=\Theta_j}^{N} P_n(\vec{\Theta})$, is a function of the limiting probability distribution. An alternate, and equivalent, definition provided in Equation 39 represents the costs incurred in each state n. In state $n < N$, the penalty incurred per unit of time is $\sum_{j=1}^{m} k_j$ such that $\Theta_m \leq n \leq \Theta_{m+1} - 1$. Since $\Theta_I = N$ and

all patients are redirected when all beds are occupied, the penalty per unit time in state $n = N$ is $\sum_{j=1}^{I} k_j$.

$$K(\vec{\Theta}) = \sum_{j=1}^{I} k_j \sum_{n=\Theta_j}^{N} P_n(\vec{\Theta}) \tag{38}$$

$$= \sum_{m=1}^{I-1} \left(\sum_{j=1}^{m} k_j \right) \left(\sum_{n=\Theta_m}^{\Theta_{m+1}-1} P_n(\vec{\Theta}) \right) + \sum_{j=1}^{I} k_j P_N(\vec{\Theta}) \tag{39}$$

In addition to the objective of minimizing the average cost, hospitals seek to minimize the percentage of the time for which all beds are in use and there is no room for incoming patients. While we examine the redirection of patients due to the threshold levels in our definition of average cost, here we reserve the phrase diversion to correspond to the condition under which all beds are full and all patients must be redirected, regardless of the patient type.

Definition Diversion rate pertains to the percentage of time that a hospital unit is completely full and must go "on diversion," redirecting all patients regardless of type. Given a redirection policy $\vec{\Theta}$ the diversion rate is $P_N(\vec{\Theta})$.

The final objective we consider is to maximize the average utilization in alignment with the hospital's goal of using resources efficiently.

for a hospital to efficiently utilize resources, specifically beds.

Definition Average utilization pertains to the percentage of beds that are utilized on average with the use of a particular redirection policy. Given a redirection policy $\vec{\Theta}$, the average utilization is $U(\vec{\Theta}) = \frac{\sum_{n=0}^{N} n\, P_n(\vec{\Theta})}{N}$

4.4 Structural Properties

In this section, we define structural properties for the problem of identifying optimal thresholds for a redirection policy. In Section 4.4.1, we define the impact of an

incremental decrease in one threshold value on the three objectives discussed above. Consequently, we characterize the properties necessary for a redirection policy such that a decrease in any of the threshold values causes an increase in the average cost.

Similarly, in Section 4.4.2 we define the impact of an incremental increase in one threshold value on the objectives defined above. Consequently, we identify the necessary properties such that an incremental increase in any of the threshold values causes an increase in the average cost.

Utilizing these properties, we then define necessary conditions for an optimal redirection policy that minimizes the average cost in Section 4.4.3. These necessary conditions can also be used to define a locally optimal redirection policy.

4.4.1 Impact of Incremental Decreases to Threshold Values

We seek to show how an incremental decrease in a threshold value impacts the performance metrics average cost, average utilization, and diversion rate. We examine the impact by comparing two policies which are equivalent except that the threshold for patient type i in policy $\vec{\Theta}$ is Θ_i and $\Theta_i^0 = \Theta_i - 1$ in policy $\vec{\Theta}^0$ (see Equations 40 and 41).

$$\vec{\Theta} = (\Theta_1, \ldots, \Theta_{i-1}, \Theta_i, \Theta_{i+1}, \ldots, \Theta_I) \tag{40}$$

$$\vec{\Theta}^0 = (\Theta_1, \ldots, \Theta_{i-1}, \Theta_i - 1, \Theta_{i+1}, \ldots, \Theta_I) \tag{41}$$

It follows from the preference definitions that $\vec{\Theta}^0$, as defined in Equation 41, is a valid, and feasible, redirection policy, satisfying the preference constraints, if and only if $\Theta_{i-1} < \Theta_i$.

We first examine the impact of an incremental decrease in the threshold for some patient i on the limiting probability distribution. Let γ_i be the ratio between the arrival rates associated with thresholds for patients $i+1$ and i (Equation 42). We define $A(i, \vec{\Theta})$ as the total limiting probability of being in a state with at most $\Theta_i - 1$ patients in the unit with admission control policy $\vec{\Theta}$ (Equation 43). $B(i, \vec{\Theta})$ is the

sum of the costs from being in states with at most $\Theta_i - 1$ patients with the admission control policy $\vec{\Theta}$ (Equation 44).

$$\gamma_i = \frac{\rho_{i+1}}{\rho_i} \tag{42}$$

$$A(i, \vec{\Theta}) = \sum_{k=0}^{\Theta_i - 1} P_k(\vec{\Theta}) \tag{43}$$

$$B(i, \vec{\Theta}) = \sum_{m=1}^{i-1} (\sum_{j=1}^{m} k_j)(\sum_{n=\Theta_m}^{\Theta_{m+1}-1} P_n(\vec{\Theta})) \tag{44}$$

Lemma 4.4.1. *Assuming a feasible patient redirection policy $\vec{\Theta}$ (Equation 40) such that $\Theta_{i-1} < \Theta_i$, then the limiting probability distribution for a similar policy in which the threshold for patient i is decremented by one, $\vec{\Theta}^0$ (Equation 41) is*

$$P_n(\vec{\Theta}^0) = \begin{cases} \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))} & \text{if } n \leq \Theta_i - 1 \\ \frac{\gamma_i \, P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))} & \text{if } n \geq \Theta_i \end{cases} \quad \forall n \in \{0 \dots N\} \tag{45}$$

Proof of Lemma 4.4.1 is provided in Appendix D. It follows from Lemma 4.4.1, that when the threshold for some patient i decreases the diversion rate under the new policy is $P_N(\vec{\Theta}^0) = \frac{\gamma_i P_N(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))}$. Since $\gamma_i < 1$ for all i, the decrease in one threshold value always results in a decrease in the diversion rate regardless of which threshold is changed. Accordingly, a decrease in the threshold for patient type i also causes a decrease in the overall utilization. Utilizing the characteristics for the changes in the limiting probabilities, the change to the average cost as a function of $\vec{\Theta}$ can also be calculated (Lemma 4.4.2).

Lemma 4.4.2. *Assuming a feasible patient redirection policy $\vec{\Theta}$ (Equation 40) such that $\Theta_{i-1} < \Theta_i$, then the average cost for a similar policy in which the threshold for patient i is decremented by one, $\vec{\Theta}^0$ (Equation 41) is*

$$K(\vec{\Theta^0}) = \frac{B(i, \vec{\Theta}) + \gamma_i(K(\vec{\Theta}) - B(i, \vec{\Theta})) + k_i \, P_{\Theta_i - 1}(\vec{\Theta})}{A(i, \vec{\Theta}) + \gamma_i(1 - A(i, \vec{\Theta}))} \tag{46}$$

Proof. To prove this equivalence, we first define $K(\vec{\Theta^0})$. Note that when the value of the threshold for patients of type i is decreased by one, the cost associated with the limiting probability of state $\Theta_i^0 = \Theta_i - 1$, the new threshold for patients of type i, increases by k_i in alignment with the definition of the average cost function. Using the results of Lemma 4.4.1, the derivation of the average cost for the new policy is shown below.

$$K(\vec{\Theta}^0) = \sum_{m=1}^{I-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta_m^0}^{\Theta_{m+1}^0-1} P_n(\vec{\Theta^0})\right) + \sum_{j=1}^{I} k_j P_N(\vec{\Theta^0}) \tag{47}$$

$$= \sum_{m=1}^{i-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta_m^0}^{\Theta_{m+1}^0-1} P_n(\vec{\Theta^0})\right) + \sum_{m=i}^{I-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta_m^0}^{\Theta_{m+1}^0-1} P_n(\vec{\Theta^0})\right) \tag{48}$$
$$+ \sum_{j=1}^{I} k_j P_N(\vec{\Theta}^0)$$

$$= \sum_{m=1}^{i-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta_m}^{\Theta_{m+1}-1} \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))}\right) \tag{49}$$
$$+ k_i \frac{P_{\Theta_i - 1}}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))}$$
$$+ \sum_{m=i}^{I-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta_m}^{\Theta_{m+1}-1} \frac{\gamma_i P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))}\right)$$
$$+ \sum_{j=1}^{N} \frac{k_j \gamma_i P_N(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))}$$

$$= \frac{B(i,\vec{\Theta}) + \gamma_i(K(\vec{\Theta}) - B(i,\vec{\Theta})) + k_i\, P_{\Theta_i - 1}(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))} \tag{50}$$

□

We use the findings from Lemma 4.4.2 to define sufficient conditions for an optimal policy that minimizes cost in Section 4.4.3.

4.4.2 Impact of Incremental Increases to Threshold Values

Similar to above, we define the impact of an incremental change to one threshold value on system performance. We seek to show how an incremental increase in a threshold value impacts the performance metrics average cost, average utilization, and diversion rate. We illustrate the impact by comparing two policies which are equivalent except that the threshold for patients i in policy $\vec{\Theta}$ is Θ_i and in policy $\vec{\Theta}'$ is $\Theta'_i = \Theta_i + 1$. Thus, the threshold for patients i in policy $\vec{\Theta}'$ is one more than the threshold for patients of type i in policy $\vec{\Theta}$. Definitions of these policies are provided in Equations 51 and 52.

$$\vec{\Theta} = (\Theta_1, \ldots, \Theta_{i-1}, \Theta_i, \Theta_{i+1}, \ldots, \Theta_I) \tag{51}$$

$$\vec{\Theta}' = (\Theta_1, \ldots, \Theta_{i-1}, \Theta_i + 1, \Theta_{i+1}, \ldots, \Theta_I) \tag{52}$$

Corresponding to preference constraints, $\vec{\Theta}'$ is a valid redirection policy if and only if $\Theta_i < \Theta_{i+1}$.

In Lemmas 4.4.3 and 4.4.4, we define the impact of this incremental increase to the threshold for patients of type i on the limiting probability distributions and the average cost. We utilize the notation for γ_i, $A(i, \vec{\Theta})$, and $B(i, \vec{\Theta})$ defined in the previous section.

Lemma 4.4.3. *Assuming a feasible patient redirection policy $\vec{\Theta}$ (Equation 51) such that $\Theta_i < \Theta_{i+1}$, then the limiting probability distribution for a similar policy in which the threshold for patient i is increased by one, $\vec{\Theta}'$ (Equation 52) is*

$$P_n(\vec{\Theta}') = \begin{cases} \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) + \frac{1}{\gamma_i}(1 - A(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}))} & \text{if } n \leq \Theta_i \\ \frac{\frac{1}{\gamma_i} P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) + \frac{1}{\gamma_i}(1 - A(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}))} & \text{if } n > \Theta_i \end{cases} \quad \forall n \in 0 \ldots N \tag{53}$$

Proof of Lemma 4.4.3 is provided in Appendix D. In contrast to the results when a threshold value is decreased by one, when a threshold value is increased by one the

average utilization and the diversion rate both increase. An increase in a threshold value may cause either an increase or a decrease in the average cost. Utilizing the characteristics for changes in limiting probabilities due to an incremental increase in the threshold for patients of type i, we define the average cost with the new control policy, $\vec{\Theta}'$ in Lemma 4.4.4.

Lemma 4.4.4. *Assuming a feasible patient redirection policy $\vec{\Theta}$ (Equation 51) such that $\Theta_i < \Theta_{i+1}$, then the average cost for a similar policy in which the threshold for patient i is increased by one, $\vec{\Theta}'$ (Equation 52) is*

$$K(\vec{\Theta}') = \frac{B(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) \sum_{j=1}^{i} k_j + \frac{1}{\gamma_i}(K(\vec{\Theta}) - B(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}) \sum_{j=1}^{i} k_j) - k_i \, P_{\Theta_i}(\vec{\Theta})}{A(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) + \frac{1}{\gamma_i}(1 - A(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}))} \tag{54}$$

Proof. To prove this equivalence, we first define $K(\vec{\Theta}')$. Note that when the value of the threshold for patients of type i is increased by one, the cost associated with the limiting probability of state Θ_i decreases by k_i in alignment with the definition of the average cost function. Using the results of Lemma 4.4.3, the average cost for the new policy is derived below.

$$K(\vec{\Theta}') = \sum_{m=1}^{I-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta'_m}^{\Theta'_{m+1}-1} P_n(\vec{\Theta}')\right) + \sum_{j=1}^{I} k_j P_N(\vec{\Theta}') \tag{55}$$

$$= \sum_{m=1}^{i-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta'_m}^{\Theta'_{m+1}-1} P_n(\vec{\Theta}')\right) + \sum_{m=i}^{I-1}(\sum_{j=1}^{m} k_j)\left(\sum_{n=\Theta'_m}^{\Theta'_{m+1}-1} P_n(\vec{\Theta}')\right) \tag{56}$$

$$+ \sum_{j=1}^{I} k_j P_N(\vec{\Theta}')$$

$$= \frac{B(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta})\sum_{j=1}^{i} k_j + \frac{1}{\gamma_i}(K(\vec{\Theta}) - B(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta})\sum_{j=1}^{i} k_j) - k_i\, P_{\Theta_i}(\vec{\Theta})}{A(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) + \frac{1}{\gamma_i}(1 - A(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}))} \tag{57}$$

□

Utilizing the properties from Lemma 4.4.4, we infer the necessary characteristics of a threshold policy such that an incremental increase in the value in one threshold (i) causes an increase in the average total cost.

4.4.3 Necessary Conditions for a Local Minimum

Since any incremental increase in utilization results in an increase in the diversion rate there are no strictly dominating policies with respect to all three objectives. As a result, we focus on identifying policies in which the average cost is minimized.

A policy $\vec{\Theta}$ is a local minimum with respect to the average cost, if all feasible incremental changes to any of the threshold values do not further decrease the average cost. Theorem 4.4.5 defines the necessary conditions for a policy $\vec{\Theta}$ to be a local or global minimum with respect to average cost.

Theorem 4.4.5. *For any optimal control policy $\vec{\Theta}$ that minimizes the average cost,*

the criteria in Equations 58 and 59 hold.

$$K(\vec{\Theta}) \leq \frac{B(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) \sum_{j=1}^{i} k_j + \frac{1}{\gamma_i}(K(\vec{\Theta}) - B(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}) \sum_{j=1}^{i} k_j) - k_i\, P_{\Theta_i}(\vec{\Theta})}{A(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) + \frac{1}{\gamma_i}(1 - A(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta}))}$$

$$\forall i \in \{1 \dots I-1\} : \Theta_i < \Theta_{i+1} \tag{58}$$

$$K(\vec{\Theta}) \leq \frac{B(i,\vec{\Theta}) + \gamma_i(K(\vec{\Theta}) - B(i,\vec{\Theta})) + k_i\, P_{\Theta_i - 1}(\vec{\Theta})}{A(i,\vec{\Theta}) + \gamma_i(1 - A(i,\vec{\Theta}))}$$

$$\forall i \in \{1 \dots I-1\} : \Theta_{i-1} < \Theta_i, \Theta_0 = 0 \tag{59}$$

Proof. Assume to the contrary that Equation 58 does not hold for some i, then there exists a policy such that increasing the threshold for patients of type i results in a lower average cost (Lemma 4.4.4), thus implying that policy $\vec{\Theta}$ is not an optimal solution.

Assume to the contrary that Equation 59 does not hold for some i, then there exists a policy such that decreasing the threshold for patients of type i results in a lower average cost (Lemma 4.4.2), thus implying that policy $\vec{\Theta}$ is not an optimal solution. □

We conjecture that if the criteria in Equation 58 and 59 hold for some $\vec{\Theta}$, this implies that $\vec{\Theta}$ is the global optimal solution for minimizing the average cost in this patient redirection problem with preference constraints.

4.5 Analysis

In addition to the identification of structural properties for the patient redirection problem with preference constraints, we utilize a numeric example to illustrate that the criteria listed in Theorem 4.4.5 hold for the optimal policy in our example. The optimal policy is identified through calculation of the limiting probability distribution, average cost, diversion rate, and average utilization for each of the 286 feasible policies, using the equations defined above (Equations 36, 37, 38) . We present an example

such that the optimal policy allows for incremental increases and decreases to all thresholds, to further demonstrate the accuracy of our claim. As a result, the example we present has a high ratio of arrivals to departures, representing a severely overloaded system. The definition of the instance for this example is provided in Section 4.5.1. We show that the necessary conditions for optimality are met by the optimal control policy for this example in Section 4.5.2.

With this example, we utilize the results from the enumeration of all feasible policies to explore the following research questions:

- How do the preference constraints for patient types impact the system performance with respect to the three objectives? (Section 4.5.3)
- What is the nature of the tradeoffs between average cost, average utilization, and rate of diversion with respect to patient redirection policies? (Section 4.5.4)

4.5.1 Instance Definition

We utilize a small example, in which the performance metrics for all feasible policies can be enumerated to instruct this study. This example assumes a unit with 10 beds and four patient types. The service rate (μ) for all patients, regardless of type, is 0.15. Parameters for the arrival rates and costs associated with each of the patient types are provided in Table 17.

Table 17: Patient specific parameters for numeric analysis

Patient Type (i)	Arrival Rate (λ_i)	Cost (k_i)
1	0.6	1
2	0.5	1
3	0.75	3
4	0.5	5

The optimal policy with respect to average cost for the instance defined above, as identified by enumeration of all 286 feasible policies, is $\vec{\Theta} = (5, 6, 9, 10)$. The discrepancies in the threshold levels for the different patient types is partially driven

by the high ratio between arrivals and departures and the small size of the hospital unit. For the optimal policy, the average cost is minimized at 2.9773, with an average utilization of 0.7343, and a diversion rate of 0.0673.

4.5.2 Necessary Conditions for Optimality

With this example, we demonstrate that the necessary optimality conditions listed in Theorem 4.4.5 hold for the policy $\vec{\Theta} = (5, 6, 9, 10)$. We achieve this by showing that the right hand side (RHS) of Equations 58 and 59 are greater or equal to $K(\vec{\Theta})$. The results from this comparison are provided in Table 18.

Table 18: Verifying necessary optimality conditions for policy $\vec{\Theta} = (5, 6, 9, 10)$

$K(\vec{\Theta})$	i	Θ_i	γ_i	k_i	$A(i, \vec{\Theta})$	$B(i, \vec{\Theta})$	$P_{\Theta_{i-1}}(\vec{\Theta})$	$P_{\Theta_i}(\vec{\Theta})$	$P_{\Theta_i}(\vec{\Theta}) \sum_{j=1}^{i} k_j$	RHS Eq. 58	RHS Eq. 59
2.97734	1	5	0.745	1	0.03780	0	0.02884	0.09037	0.09037	2.98465	2.97747
2.97734	2	6	0.714	1	0.12817	0.09037	0.09037	0.17572	0.35144	2.98475	2.98687
2.97734	3	9	0.4	3	0.73108	1.29599	0.21790	0.20176	1.0088	3.07105	3.12696

The necessary optimal conditions needed for a policy that minimizes cost was tested on a variety of other instances, in which enumeration of all policies allowed for the identification of the optimal solution. For all these experiments, the globally optimal policy was the only policy for which the necessary optimality conditions held, implying that there were no policies that were local minimums and not globally optimal.

4.5.3 Impact of Preferences

One of the key assumptions of our model is that preferences for the redirection of patients are fixed and consistent with the redirection costs k_i, regardless of the arrival rates for each patient type. Utilizing the same example as above, we examine the impact of the preference constraints by developing three additional instances with rearranged preferences.

When rearranging the preferences for patient types, we assume that costs associated with each threshold in the original setting remain fixed, and as a result through

the changes to the preferences, the costs associated with each patient type change. The arrival rates for the patient types remain the same. In Table 19, we present the performance metrics for the policies when average cost is minimized in the four instances representing variations of patient preferences.

Table 19: Impact of changes to patient type preferences in four-threshold policies on average cost, utilization, and diversion rate

Instance	Patient Type Assignments to Thresholds				Average Cost	Utilization	Diversion Rate	Optimal Policy
	Θ_1 ($k_1 = 1$)	Θ_2 ($k_2 = 1$)	Θ_3 ($k_3 = 3$)	Θ_4 ($k_4 = 5$)				
1	1 $\lambda = .6$	2 $\lambda = .5$	3 $\lambda = .75$	4 $\lambda = .5$	2.9773	0.7343	0.0673	(5, 6, 9, 10)
2	3 $\lambda = .75$	1 $\lambda = .6$	2 $\lambda = .5$	4 $\lambda = .5$	2.4630	0.6975	0.0846	(4, 6, 10, 10)
3	4 $\lambda = .5$	2 $\lambda = .5$	3 $\lambda = .6$	1 $\lambda = .75$	3.2510	0.7600	0.0918	(6, 6, 9, 10)
4	4 $\lambda = .5$	2 $\lambda = .5$	1 $\lambda = .75$	3 $\lambda = .6$	3.3273	0.7673	0.1753	(5, 5, 10, 10)

For this example, a system in which the lowest cost patients arrive with the greatest frequency (Instance 2) results in the lowest cost solution, accompanied by the lowest utilization. When, instead, the patients with the highest cost have the highest arrival rates (Instance 3), a balance of the three objectives is achieved. A similar balanced optimal policy occurs for Instance 1, in which the arrival rates are not strictly increasing or decreasing with the thresholds. The optimal policy in Instance 1 achieves the lowest diversion rate and second lowest average cost.

These results illustrate that the assumption of preference constraints can significantly impact how well the system performs and the best admission control policy with respect to average cost. In a hospital setting, the appropriateness of the hospital unit for the patient type can not be disregarded, and patient redirection policies must be made within this context. It is apparent that in this example, the proportion of patients in each group is a key consideration in the definition of a redirection policy. Additional experimentation is needed to examine the impact of preferences for other instances and variations on patient preferences.

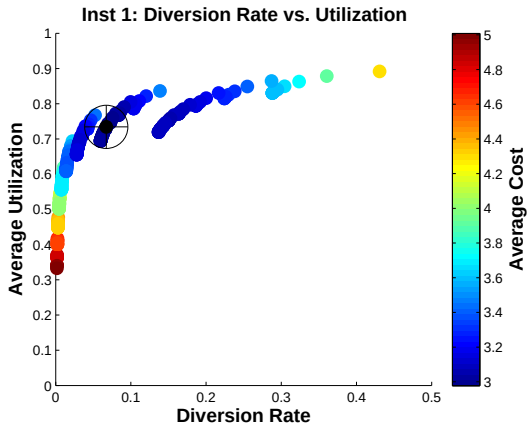
4.5.4 Tradeoffs in Average Cost, Diversion Rate, and Utilization

While the optimal policy with respect to average cost is identified, for each feasible policy there is an intrinsic tradeoff in all three objectives: average cost, average utilization, and diversion rate. To examine these tradeoffs we enumerate all feasible threshold policies for the four instances defined in Table 19. For each policy we identify the limiting probability distribution and correspondingly calculate the average cost, diversion rate, and utilization with the equations provided above (Equations 36, 37, 38).

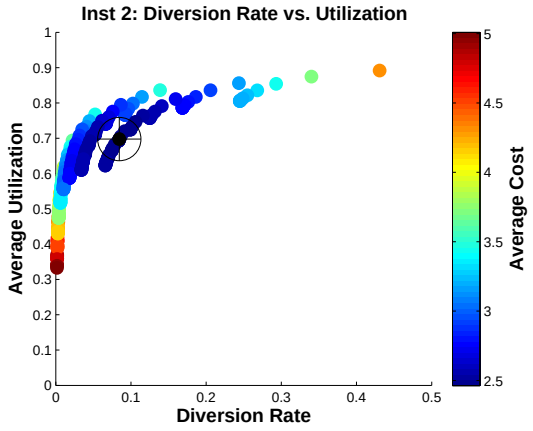
In Figure 17, for each instance we plot the associated diversion rate and average utilization for each feasible policy, represented as a point on the graph. The color of each point corresponds to the average cost. This allows for an examination of the tradeoffs that exist in the three objectives. The policy with the minimum average cost, among all instances, is highlighted with a circle on the graph.

For all four instances the minimum average cost policy coincides with the portion of the graph where the relative costs of improving the diversion rate and improving average utilization are balanced. While the minimum cost policy appears to balance average utilization and the diversion rate, it does not achieve the best solution with respect to these objectives. Instead, for all four instances a simultaneous increase in utilization and decrease in the diversion rate can be achieved, although it is accompanied by an increase in the average cost. Thus, for these four instances, utilizing the objective of minimizing the average cost, the resulting optimal policy simultaneously balances the tradeoffs between utilization and the diversion rate, although it does not produce the optimal solution with respect to either objective. Thus, through the use of the objective to minimize average cost we identify a policy on the efficient frontier in which all three of the objectives are accounted for.

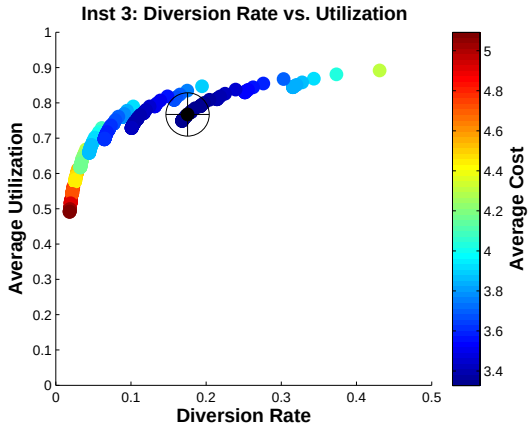
Figure 17: Tradeoff in the diversion rate, utilization, and average cost for all feasible policies for Instances 1, 2, 3, and 4 with a designation of the minimum average cost policy



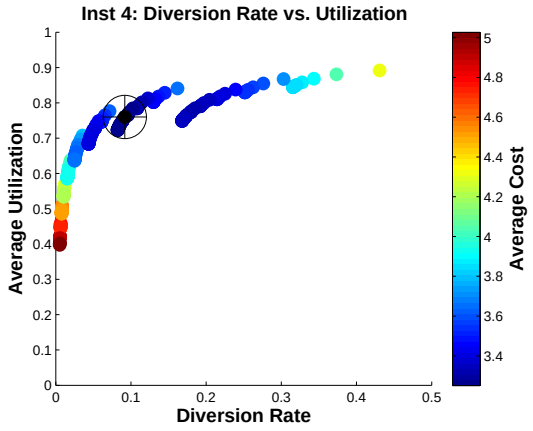
(a) Instance 1



(b) Instance 2



(c) Instance 3



(d) Instance 4

4.6 Conclusion

As previously discussed, the example used in the computational analysis has a high ratio of demand by patients to supply of beds in order to demonstrate the accuracy of the necessary conditions for a policy that minimizes average cost. We illustrate that the necessary conditions hold for the globally optimal solution in the example (see Section 4.5.2. Additional experimentation, not reported here, showed similar results for other instances, leading us to conjecture that these conditions may be sufficient for identifying a globally optimal solution.

In the four instances we examine, we show that the use of the objective of minimizing average cost simultaneously provides a solution that balances tradeoffs in average utilization and the diversion rate. Additionally, with the assumption of preference constraints and patient dependent costs, arrival rates significantly impact the optimal control policy, or patient redirection policy. Additional experimentation with other instances is needed to further verify these results and to conduct an analysis of the sensitivity of the tradeoffs in the objectives to changes to arrival rates and service times. Specifically, a simulation approach is needed to study the impact of the assumption of uniform service rates for all patient types.

In addition to the numeric examples, we have provided the first model to our knowledge that examines the development of a multi-threshold patient redirection policy with the assumption of preference constraints for redirection of patients. With this model, we define necessary conditions for a threshold policy that minimizes the average cost. These characteristics are especially useful for circumstances in which enumeration of all feasible policies may not be possible or ideal. If our conjecture that these characteristics are both necessary and sufficient to define global optimality is proven, this could inform the development of algorithms to efficiently identify the optimal policy in large instance settings.

While we have made a contribution in defining the problem of modeling multi-threshold patient redirection policies with preference constraints, there exist opportunities to expand on this work. First, we seek to prove or provide a counterexample for the conjecture regarding sufficient conditions for the global optimal policy. Additionally, with the assumption that these policies will be implemented by hospitals that are part of a larger group of hospitals, simulation can be used to examine the interactions between the definition of threshold policies at associated hospitals. Additionally, by focusing on these interactions in a centralized environment, an examination of how policies that only account for local unit census information can be improved or how

control policies that simultaneously account for the census at all system hospitals should be examined. The results from these extensions to our model may provide a better understanding of the multi-threshold patient redirection problem and aid in the implementation of these policies in hospitals.

APPENDIX A

AFRICAN HEALTH CARE ALLOCATION MODEL FORMULATION

Sets:

A	=	set of arcs in the network of health care facilities incorporating the hierarchical distribution system
H	=	set of health centers
H'	=	set of primary-level health centers, H' $\subset$ H
T	=	set of treatments
P	=	set of population centers

Data:

$N_{p,t}$ = number of people in need of treatment t from population center p

$V_{p,t}$ = increase in health value (in QALYs) from providing treatment t to an individual from population center p

$R_{p,h}$ = "traveling cost" from population center p to health center h

$D_{m,n,t}$ = "shipping cost" of a treatment t from health center m to health center n

I_t = cost of a treatment t

B = total budget for the "distribution cost"

O_h = fixed-cost of opening health center h

Y_h = number of nurse-days available at health center h

Q_t = time to deliver treatment t to one patient

L = time available per nurse-day

$$C_{p,g,h} = \begin{cases} 1 & \text{if the cost of travel between population center } p \text{ and health center } h \\ & \text{is less than the cost of travel between population center } p \text{ and health center } g \\ 0 & \text{otherwise} \end{cases}$$

Variables:

$a_{h,t}$ = quantity of treatments t that are allocated to health center h

$r_{p,h,t}$ = number of individuals from population center p requiring treatment that successfully receive treatment at health center h

$q_{m,n,t}$ = quantity of treatment t transported from health center m to health center n

$k_{p,h}$ = number of nurse-days at health center h assigned to population p

$$y_h = \begin{cases} 1 & \text{if health center } h \text{ is opened} \\ 0 & \text{otherwise} \end{cases}$$

$$t_{p,h} = \begin{cases} 1 & \text{if } h \text{ is the closest open health center which provides treatment to } p \\ 0 & \text{otherwise} \end{cases}$$

Constraints:

$$\sum_{\substack{(f,g)\in A\\ t\in T}} q_{f,g,t}D_{f,g,t} + \sum_{h\in H, t\in T} I_t a_{h,t} + \sum_{h\in H} O_h y_h \leq B \quad \forall f \in H, g \in H \tag{60}$$

$$\sum_{(f,g)\in A} q_{f,g,t} + a_{f,t} = \sum_{(e,f)\in A} q_{e,f,t} \quad \forall f \in H\ H', t \in T \tag{61}$$

$$\sum_{p\in P} r_{p,h,t} \leq a_{t,h} \quad \forall h \in H, t \in T \tag{62}$$

$$r_{p,h,t} \leq N_{p,t} t_{p,h} \quad \forall h \in H, p \in P, t \in T \tag{63}$$

$$\sum_{h\in H} t_{p,h} = 1 \quad \forall p \in P \tag{64}$$

$$t_{p,g} \leq 1 - \sum_{h\in H} \frac{C_{p,g,h} y_h}{|H|} \quad \forall p \in P, g \in H \tag{65}$$

$$t_{p,h} \leq y_h \quad \forall p \in P, h \in H \tag{66}$$

$$\sum_{p\in P} N_{p,t} y_h \geq a_{h,t} \quad \forall h \in H, t \in T \tag{67}$$

$$\sum_{p\in P} k_{p,h} \leq Y_h \quad \forall h \in H \tag{68}$$

$$k_{p,h}(L - R_{p,h}) \geq \sum_{t\in T} r_{p,h,t} Q_t \quad \forall p \in P, h \in H \tag{69}$$

$$k_{p,h} \leq t_{p,h} Y_h \quad \forall p \in P, h \in H \tag{70}$$

$$k_{p,h} \geq 0 \quad \forall p \in P, h \in H \tag{71}$$

$$r_{p,h,t} \geq 0 \quad \forall p \in P, h \in H, t \in T \tag{72}$$

$$a_{t,h}, r_{p,t,h}, q_{m,n,t}, k_{p,h} \in \mathbb{Z} \tag{73}$$

$$y_h, t_{p,h} \in \mathbb{B} \tag{74}$$

Objectives:

$$max \quad \sum_{p\in P, h\in H, t\in T} V_{p,t} r_{p,h,t} \tag{75}$$

$$min \quad \sum_{\substack{(f,g)\in A\\ t\in T}} q_{f,g} D_{f,g,t} + \sum_{t\in T, h\in H} I_t a_{h,t} + \sum_{h\in H} O_h y_h \tag{76}$$

$$max \quad min_{p\in P} \frac{\sum_{t\in T, h\in H} r_{p,h,t} V_{p,t}}{\sum_{t\in T} N_{p,t} V_{p,t}} \tag{77}$$

$$min \quad max_{p\in P} \frac{\sum_{t\in T, h\in H} r_{p,h,t} V_{p,t}}{\sum_{t\in T} N_{p,t} V_{p,t}} - min_{p\in P} \frac{\sum_{t\in T, h\in H} r_{p,h,t} V_{p,t}}{\sum_{t\in T} N_{p,t} V_{p,t}} \tag{78}$$

The first constraint (60) defines restrictions on the total cost, including costs for opening of health centers, distribution, and procurement, with regard to the budget. The second constraint (61) depicts flow constraints in the distribution and allocation of treatments to health centers. This constraint ensures that treatments are always distributed from regional health centers to rural health centers. Limits on the number of individuals treated at a health center due to supply and demand, are defined in constraints (62) and (63), respectively. Constraints (64),(65), and (66) ensure that each population center is only served by nurses from the closest health center. The next constraint (67) states that treatments can only be allocated to open health centers. The number of available nurse-days is incorporated in constraint (69). Constraints (70) and (71) describe the correlation between the distance traveled by a nurse to a population center and the maximum number of individuals that can be treated by the nurse. The last two constraints concern the nonnegativity of the number of nurses dispatched from a health center and the number of patients receiving treatment at a population center.

The first and second objectives maximize effectiveness and minimize total costs, respectively. The third and fourth objectives relate to the equity measures described above. The third objective maximizes the smallest success rate among all population centers. The fourth objective minimizes the difference between the highest and lowest success rates in all communities.

APPENDIX B

BED ASSIGNMENT INTEGER PROGRAM MODEL FORMULATION

Sets:

$P =$ set of patients

$E =$ set of events, or decision epochs, including patient arrivals, service completions, and exits; $|E| = 3|P|$

$U =$ set of downstream units, $\{1 \dots \mathrm{J}\}$

Variables:

$$v_{p,j,e} = \begin{cases} 1 & \text{if patient } p \text{ is in service in unit } j \text{ at time of event } e \\ 0 & \text{otherwise} \end{cases}$$

$$x_{p,j,e} = \begin{cases} 1 & \text{if patient } p \text{ is assigned to unit } j \text{ at event } e \text{ (i.e., patient } p \text{ begins service in unit } j \text{ at event } e) \\ 0 & \text{otherwise} \end{cases}$$

$$w_p = \begin{cases} 1 & \text{if patient } p \text{ is accepted to the upstream unit upon arrival} \\ 0 & \text{otherwise} \end{cases}$$

$$u_p = \begin{cases} 1 & \text{if patient } p \text{ exits the system while boarding, and is never assigned to a downstream unit} \\ 0 & \text{otherwise} \end{cases}$$

(79)

Data:

$$R_{p,e} = \begin{cases} 1 & \text{if event } e \text{ corresponds to the arrival of patient } p \\ 0 & \text{otherwise} \end{cases}$$

$$C_{p,e} = \begin{cases} 1 & \text{if event } e \text{ corresponds to the service completion of patient } p \\ 0 & \text{otherwise} \end{cases}$$

$$X_{p,e} = \begin{cases} 1 & \text{if event } e \text{ corresponds to the exit of patient } p \\ 0 & \text{otherwise} \end{cases}$$

$E_p^R =$ Event corresponding to arrival of patient p

$E_p^C =$ Event corresponding to service completion of patient p

$E_p^X =$ Event corresponding to exit of patient p

$T_p^C =$ Time of the arrival of patient p

$T_p^X =$ Time of the exit of patient p

$T_e =$ Time of event e

$\Delta_e =$ Time between events e and $e+1$

$b_p =$ Cost of boarding of patient p for one unit of time

$a_{p,j} =$ Value of patient p being served in unit j for one unit of time

$N_0 =$ Capacity of upstream unit

$N_j =$ Capacity of downstream unit j

Objective and Constraints:

$$\max \sum_{p\in P}\sum_{j=1}^{J}\sum_{e\in E} v_{p,j,e}(a_{p,j}\,\Delta_e) - x_{p,j,e}\,b_p\,(T_e - T_p^C) - u_p\,b_p(T_p^X - T_p^C) \tag{80}$$

$$\text{s.t.} N_0 - w_m\,R_{m,e} + \sum_{e=1}^{E_p^R-1}\sum_{m\in P} u_m\,X_{m,e} + \sum_{j=1}^{J} x_{p,j,e} \le N_0\,w_p \qquad \forall p\in P \tag{81}$$

$$w_p - u_p - \sum_{e=E_p^C}^{E_p^X}\sum_{j=1}^{J} x_{p,j,e} = 0 \qquad \forall p\in P \tag{82}$$

$$\sum_{e=1}^{E_p^C}\sum_{j=1}^{J} x_{p,j,e} + \sum_{e=E_p^X}^{|E|}\sum_{j=1}^{J} x_{p,j,e} = 0 \qquad \forall p\in P \tag{83}$$

$$\sum_{e=0}^{f}\sum_{p\in P} w_p R_{p,e} - u_p X_{p,e} - \sum_{j=1}^{J} x_{p,j,e} \le N_0 \qquad \forall f\in E \tag{84}$$

$$\sum_{e=0}^{f} x_{p,j,e} - v_{p,j,e} \ge 0 \qquad \forall f\in E, p\in P, j\in U \tag{85}$$

$$1 - \sum_{e=0}^{f}(X_{p,e} + v_{p,j,e}) \ge 0 \qquad \forall f\in E, p\in P, j\in U \tag{86}$$

$$\sum_{e=0}^{f}(x_{p,j,e} - X_{p,e}) \ge 0 \qquad \forall f\in E, p\in P, j\in U \tag{87}$$

$$\sum_{e=0}^{f}(x_{p,j,e} - X_{p,e} - v_{p,j,e}) \ge 0 \qquad \forall f\in E, p\in P, j\in U \tag{88}$$

$$\sum_{p\in P} v_{p,j,e} \le N_j \qquad \forall e\in E, j\in U \tag{89}$$

$$v_{p,j,e}, x_{p,j,e} \ge 0 \qquad \forall f\in E, p\in P, j\in U \tag{90}$$

$$v_{p,j,e}, x_{p,j,e} \text{ integer} \qquad \forall f\in E, p\in P, j\in U \tag{91}$$

$$w_p, u_p \ge 0 \qquad \forall p\in P \tag{92}$$

$$w_p, u_p \text{ integer} \qquad \forall p\in P \tag{93}$$

Constraints 81 designate that as long as the upstream unit is not full a patient must be accepted to the system. If accepted, the patient will leave the upstream unit either through assignment to a downstream unit or by exiting from boarding (Constraint 82). Assignment to downstream units can not be made before the patient completes service upstream or after exiting the system (Constraint 83). Constraints 85, 86, 87, and 88 define the events, or times, at which a patient is in service in a particular

unit requiring that the patient has been assigned to the unit, has completed upstream service, and has not exited the system. Constraints 84 and 89 ensure that the number of patients in a unit does not exceed capacity of the upstream and downstream units, respectively. The objective (80) maximizes the sum of the reward achieved between events for each patient in service in a downstream unit and the cost of boarding for both patients that are eventually assigned to downstream units and those that exit from the upstream unit.

APPENDIX C

EXPLANATION FOR CHOICE OF WARM-UP PERIOD

To identify the appropriate length of the warm-up period, we plot the number of patients in the system over time in the simulation for numerous instances. From these graphs we identify a 15-day warm up as appropriate for the simulation analysis, since the system reaches steady state beyond this point. A sample of the plots for 6 instances is provided in Figure 18.

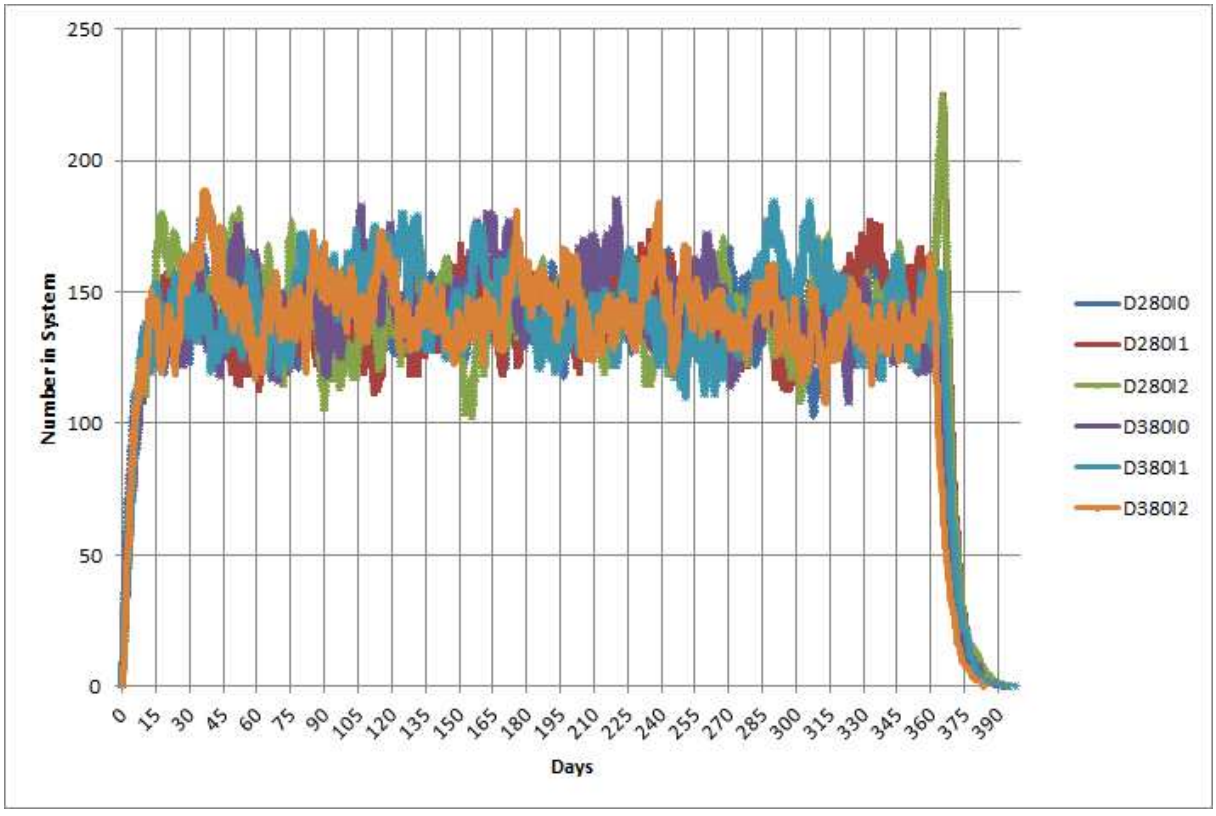


Figure 18: Plot of number of patients in system over the course of the simulation for 6 instances

APPENDIX D

PROOFS OF LEMMAS 4.4.1 AND 4.4.3

Lemma D.0.1 (Lemma 4.4.1)**.** *Assuming a feasible patient redirection policy* $\vec{\Theta}$ *(Equation 40) such that* $\Theta_{i-1} < \Theta_i$*, then the limiting probability distribution for a similar policy in which the threshold for patient i is decremented by one,* $\vec{\Theta}^0$ *(Equation 41) is*

$$P_n(\vec{\Theta}^0) = \begin{cases} \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta})+\gamma_i(1-A(i,\vec{\Theta}))} & \text{if } n \leq \Theta_i - 1 \\ \frac{\gamma_i\, P_n(\vec{\Theta})}{A(i,\vec{\Theta})+\gamma_i(1-A(i,\vec{\Theta}))} & \text{if } n \geq \Theta_i \end{cases} \quad \forall n \in \{0 \dots N\} \tag{94}$$

Proof. First, we define the limiting probabilities for policy $\vec{\Theta}^0$ in Equation 95.

$$P_n(\vec{\Theta}^0) = \begin{cases} \frac{P_0(\vec{\Theta}^0)\,{\rho_1}^n}{n!} & \forall n \in S : 0 < n \leq \Theta_1 \\ \frac{P_0(\vec{\Theta}^0)\,{\rho_1}^{\Theta_1}\,{\rho_2}^{n-\Theta_1}}{n!} & \forall n \in S : \Theta_1 < n \leq \Theta_2 \\ \dots & \dots \\ \frac{P_0(\vec{\Theta}^0)\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1} \dots {\rho_i}^{n-\Theta_{i-1}}}{n!} & \forall n \in S : \Theta_{i-1} < n \leq \Theta_i^0 = \Theta_i - 1 \\ \frac{P_0(\vec{\Theta}^0)\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1} \dots {\rho_i}^{(\Theta_i-1)-\Theta_{i-1}}\,{\rho_{i+1}}^{n-(\Theta_i-1)}}{n!} & \forall n \in S : \Theta_i - 1 = \Theta_i^0 < n \leq \Theta_{i+1} \\ \dots & \dots \\ \frac{P_0(\vec{\Theta}^0)\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1} \dots {\rho_i}^{(\Theta_i-1)-\Theta_{i+1}}\,{\rho_{i+1}}^{\Theta_{i+1}-(\Theta_i-1)} \dots {\rho_{I-1}}^{n-\Theta_{I-1}}}{n!} & \forall n \in S : \Theta_{I-1} < n \leq N \end{cases} \tag{95}$$

From this definition it follows that

$$\frac{P_n(\vec{\Theta}^0)}{P_0(\vec{\Theta}^0)} = \begin{cases} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \forall n \in S : 0 < n \leq \Theta_i^0 = \Theta_i - 1 \\ \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})}\,\frac{\rho_{i+1}}{\rho_i} = \gamma_i\,\frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \forall n \in S : \Theta_i^0 = \Theta_i - 1 < n \leq N \end{cases} \tag{96}$$

Utilizing these results, we can write the expression for $P_0(\vec{\Theta}^0)$.

$$P_0(\vec{\Theta}^0) = \frac{1}{\sum_{n=0}^{N} \frac{P_n(\vec{\Theta}^0)}{P_0(\vec{\Theta}^0)}} \tag{97}$$

$$= \frac{1}{\sum_{n=0}^{\Theta_i - 1} \frac{P_n(\vec{\Theta}^0)}{P_0(\vec{\Theta}^0)} + \sum_{n=\Theta_i}^{N} \frac{P_n(\vec{\Theta}^0)}{P_0(\vec{\Theta}^0)}} \tag{98}$$

$$= \frac{1}{\sum_{n=0}^{\Theta_i - 1} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} + \sum_{n=\Theta_i}^{N} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} \gamma_i} \tag{99}$$

$$= \frac{1}{\frac{A(i,\vec{\Theta})}{P_0(\vec{\Theta})} + \frac{1 - A(i,\vec{\Theta})}{P_0(\vec{\Theta})} \gamma_i} \tag{100}$$

$$= \frac{P_0(\vec{\Theta})}{A(i,\vec{\Theta}) + (1 - A(i,\vec{\Theta}))\,\gamma_i} \tag{101}$$

With the definition of $P_0(\vec{\Theta}^0)$ provided in Equation 101 and the relationships defined in Equation 96, we define the limiting probabilities with policy $\vec{\Theta}^0$ as a function of $\vec{\Theta}$.

$$P_n(\vec{\Theta}^0) = \begin{cases} P_0(\vec{\Theta}^0) \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} \\ = \frac{P_0(\vec{\Theta})}{A(i,\vec{\Theta}) + (1 - A(i,\vec{\Theta}))\,\gamma_i} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} \\ = \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + (1 - A(i,\vec{\Theta}))\,\gamma_i} & \forall n \in S : 0 < n \leq \Theta_i^0 = \Theta_i - 1 \\ P_0(\vec{\Theta}^0)\,\gamma_i \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} \\ = \frac{P_0(\vec{\Theta})}{A(i,\vec{\Theta}) + (1 - A(i,\vec{\Theta}))\,\gamma_i} \gamma_i \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} \\ = \frac{\gamma_i P_n(\vec{\Theta})}{A(i,\vec{\Theta}) + (1 - A(i,\vec{\Theta}))\,\gamma_i} & \forall n \in S : \Theta_i^0 = \Theta_i - 1 < n \leq N \end{cases} \tag{102}$$

□

Lemma D.0.2 (Lemma 4.4.3)**.** *Assuming a feasible patient redirection policy* $\vec{\Theta}$ *(Equation 51) such that* $\Theta_i < \Theta_{i+1}$*, then the limiting probability distribution for a similar policy in which the threshold for patient* i *is increased by one,* $\vec{\Theta}'$ *(Equation 52) is*

$$P_n(\vec{\Theta}') = \begin{cases} \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta})+P_{\Theta_i}(\vec{\Theta})+\frac{1}{\gamma_i}(1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta}))} & \text{if } n \leq \Theta_i \\ & \forall n \in 0 \dots N \\ \frac{\frac{1}{\gamma_i} P_n(\vec{\Theta})}{A(i,\vec{\Theta})+P_{\Theta_i}(\vec{\Theta})+\frac{1}{\gamma_i}(1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta}))} & \text{if } n > \Theta_i \end{cases} \tag{103}$$

Proof. This proof follows in a similar fashion as Lemma D.0.1. First, we define the limiting probabilities for policy $\vec{\Theta}'$.

$$P_n(\vec{\Theta}') = \begin{cases} \frac{P_0(\vec{\Theta}')\,{\rho_1}^n}{n!} & \forall n \in S : 0 < n \leq \Theta_1 \\ \frac{P_0(\vec{\Theta}')\,{\rho_1}^{\Theta_1}\,{\rho_2}^{n-\Theta_1}}{n!} & \forall n \in S : \Theta_1 < n \leq \Theta_2 \\ \dots & \dots \\ \frac{P_0(\vec{\Theta}')\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1}\dots{\rho_i}^{n-\Theta_{i-1}}}{n!} & \forall n \in S : \Theta_{i-1} < n \leq \Theta_i' = \Theta_i + 1 \\ \frac{P_0(\vec{\Theta}')\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1}\dots{\rho_i}^{(\Theta_i+1)-\Theta_{i-1}}\,{\rho_{i+1}}^{n-(\Theta_i+1)}}{n!} & \forall n \in S : \Theta_i + 1 = \Theta_i' < n \leq \Theta_{i+1} \\ \dots & \dots \\ \frac{P_0(\vec{\Theta}')\,{\rho_1}^{\Theta_1}\,{\rho_2}^{\Theta_2-\Theta_1}\dots{\rho_i}^{(\Theta_i+1)-\Theta_{i-1}}\,{\rho_{i+1}}^{\Theta_{i+1}-(\Theta_i+1)}\dots{\rho_{I-1}}^{n-\Theta_{I-1}}}{n!} & \forall n \in S : \Theta_{I-1} < n \leq N \end{cases} \tag{104}$$

From this definition it follows that

$$\frac{P_n(\vec{\Theta}')}{P_0(\vec{\Theta}')} = \begin{cases} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \forall n \in S : 0 < n \leq \Theta_i \\ \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})}\,\frac{\rho_i}{\rho_{i+1}} = \frac{1}{\gamma_i}\,\frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \forall n \in S : \Theta_i < n \leq N \end{cases} \tag{105}$$

Note that the multiple for Θ_i, does not change when we have an increase in the threshold for patients of type i, dissimilar from when there is a decrease in the threshold for patients of type i.

Utilizing these results, we can write the expression for $P_0(\vec{\Theta}^0)$.

$$P_0(\vec{\Theta}') = \frac{1}{\sum_{n=0}^{N} \frac{P_n(\vec{\Theta}')}{P_0(\vec{\Theta}^0)}} \tag{106}$$

$$= \frac{1}{\sum_{n=0}^{\Theta_i} \frac{P_n(\vec{\Theta}')}{P_0(\vec{\Theta}')} + \sum_{n=\Theta_i+1}^{N} \frac{P_n(\vec{\Theta}')}{P_0(\vec{\Theta}')}} \tag{107}$$

$$= \frac{1}{\sum_{n=0}^{\Theta_i-1} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} + \frac{P_{\Theta_i}(\vec{\Theta})}{P_0(\vec{\Theta})} + \sum_{n=\Theta_i+1}^{N} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} \frac{1}{\gamma_i}} \tag{108}$$

$$= \frac{1}{\frac{A(i,\vec{\Theta})+P_{\Theta_i}(\vec{\Theta})}{P_0(\vec{\Theta})} + \frac{1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta})}{P_0(\vec{\Theta})} \frac{1}{\gamma_i}} \tag{109}$$

$$= \frac{P_0(\vec{\Theta})}{A(i,\vec{\Theta}) + P_{\Theta_i}(\vec{\Theta}) + (1 - A(i,\vec{\Theta}) - P_{\Theta_i}(\vec{\Theta})) \frac{1}{\gamma_i}} \tag{110}$$

With the definition of $P_0(\vec{\Theta}')$ provided in Equation 110 and the relationships defined in Equation 105, we define the limiting probabilities with policy $\vec{\Theta}^0$ as a function of $\vec{\Theta}$.

$$P_n(\vec{\Theta}') = \begin{cases} P_0(\vec{\Theta}') \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \\ = \frac{P_0(\vec{\Theta})}{A(i,\vec{\Theta})+P_{\Theta_i}(\vec{\Theta})+(1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta}))\frac{1}{\gamma_i}} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \\ = \frac{P_n(\vec{\Theta})}{A(i,\vec{\Theta})+(1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta}))\frac{1}{\gamma_i}} & \forall n \in S : 0 < n \leq \Theta_i \\ P_0(\vec{\Theta}') \frac{1}{\gamma_i} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \\ = \frac{P_0(\vec{\Theta})}{A(i,\vec{\Theta})+P_{\Theta_i}(\vec{\Theta})+(1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta}))\frac{1}{\gamma_i}} \frac{1}{\gamma_i} \frac{P_n(\vec{\Theta})}{P_0(\vec{\Theta})} & \\ = \frac{\frac{1}{\gamma_i} P_n(\vec{\Theta})}{A(i,\vec{\Theta})+P_{\Theta_i}(\vec{\Theta})+(1-A(i,\vec{\Theta})-P_{\Theta_i}(\vec{\Theta}))\frac{1}{\gamma_i}} & \forall n \in S : \Theta_i < n \leq N \end{cases} \tag{111}$$

□

REFERENCES

[1] Akhtar, R., *Health Care Patterns and Planning in Developing Countries*. London: Greenwood Press, 1991.

[2] Alagoz, O. and Ayvaci, M., "Uniformization in Markov Decision Processes," in *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

[3] Asamoah, O. K., Weiss, S. J., Ernst, A. a., Richards, M., and Sklar, D. P., "A Novel Diversion Protocol Dramatically Reduces Diversion Hours," *The American Journal of Emergency Medicine*, vol. 26, pp. 670–675, July 2008.

[4] Ayvaz, N. and Huh, W. T., "Allocation of hospital capacity to multiple types of patients," *Journal of Revenue and Pricing Management*, vol. 9, pp. 386–398, Sept. 2010.

[5] Bahmani, B. and Kapralov, M., "Improved bounds for online stochastic matching," *Algorithms ESA 2010*, pp. 170–181, 2010.

[6] Balaji, R. and Brownlee, M., "Bed Management Optimization," tech. rep., Infosys Technologgies, 2009.

[7] Begley, C. E., Lairson, D. R., and Balkrishnan, R., *Evaluating the Healthcare System: Effectiveness, Efficiency, and Equity*. Chicago, IL: Health Administration Press, third ed., 2004.

[8] Berman, P., "Health Sector Reform in Developing Countries: Making Health Development Sustainable," *Health Policy*, vol. 32, no. 1, pp. 13–28, 1995.

[9] Boaden, R., Proudlove, N., and Wilson, M., "An exploratory study of bed management," *Journal of Management in Medicine*, vol. 13, pp. 234–50, Jan. 1999.

[10] Boucherie, R., "On the arrival theorem for product form queueing networks with blocking," *Performance Evaluation*, vol. 29, pp. 155–176, Apr. 1997.

[11] Bretthauer, K. M., Heese, H. S., Pun, H., and Coe, E., "Blocking in Healthcare Operations : A New Heuristic and an Application," *Production and Operations Management*, vol. 20, no. 3, pp. 375–391, 2011.

[12] Buchbinder, N., Jain, K., and Naor, J. S., "Online Primal-Dual Algorithms for Maximizing Ad-Auctions Revenue," in *Algorithms ESA 2007* (Arge, L., Hoffmann, M., and Welzl, E., eds.), vol. 4698, pp. 253–264, Springer, 2007.

[13] Carr, B. G., Hollander, J. E., Baxt, W. G., Datner, E. M., and Pines, J. M., "Trends in Boarding of Admitted Patients in US Emergency Departments 2003-2005," *The Journal of Emergency Medicine*, vol. 39, pp. 506–511, Oct. 2010.

[14] Carr, C. C. and Jallah, J. D., *Improving Spatial Accessibility to Antiretroviral Treatments for HIV / AIDS*. PhD thesis, University of Zaragoza, 2008.

[15] Chen, L., Evans, T., Anand, S., Boufford, J. I., Brown, H., Chowdhury, M., Cueto, M., Dare, L., Dussault, G., Elzinga, G., Fee, E., Habte, D., Hanvoravongchai, P., Jacobs, M., Kurowski, C., Michael, S., Pablos-Mendez, A., Sewankambo, N., Solimano, G., Stilwell, B., de Waal, A., and Wibulpolprasert, S., "Human Resources for Health: Overcoming the Crisis," *The Lancet*, vol. 364, no. 9449, pp. 1984–1990, 2004.

[16] Cole, R. and de Blij, H. J., *Survey of Subsaharan Africa: A Regional Geography*. New York, New York, USA: Oxford University Press, 2006.

[17] Committee on Quality of Health Care in America Institute of Medicine, *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, D.C: The National Academies Press, 2001.

[18] Csaba, B. and Pluhár, A., "A randomized algorithm for the on-line weighted bipartite matching problem," *Journal of Scheduling*, vol. 11, no. 6, pp. 449–455, 2008.

[19] Deo, S. and Gurvich, I., "Centralized vs. Decentralized Ambulance Diversion: A Network Perspective," *Management Science*, vol. 57, no. 7, pp. 1300–1319, 2011.

[20] Esogbue, A. and Singh, A. J., "A stochastic model for an optimal priority bed distribution problem in a hospital ward," *Operations Research*, vol. 24, no. 5, pp. 884–898, 1976.

[21] Falvo, T., Grove, L., Stachura, R., Vega, D., Stike, R., Schlenker, M., and Zirkin, W., "The opportunity loss of boarding admitted patients in the emergency department.," *Academic Emergency Medicine : Official Journal of the Society for Academic Emergency Medicine*, vol. 14, pp. 332–7, Apr. 2007.

[22] Feldman, J., Mehta, A., Mirrokni, V., and Muthukrishnan, S., "Online stochastic matching: Beating 1-1/e," *FOCS '09. 50th Annual IEEE Symposium on Foundations of Computer Science*, pp. 117–126, 2009.

[23] Flessa, S., "Where Efficiency Saves Lives: A Linear Programme for the Optimal Allocation of Health Care Resources in Developing Countries," *Health Care Management Science*, vol. 3, pp. 249–267, 2000.

[24] Foster, S. D., "Improving the Supply and Use of Essential Drugs in Sub-Saharan Africa," *Policy Research Working Paper Series 456, The World Bank*, 1990.

[25] Fuchs, B., Hochstattler, W., and Kern, W., "Online matching on a line," *Theoretical Computer Science*, vol. 332, pp. 251–264, Feb. 2005.

[26] Gans, N. and Van Ryzin, G., "Optimal Control of a Multiclass, Flexible Queueing System," *Operations Research*, vol. 45, no. 5, pp. 677–693, 1997.

[27] Gómez-Corral, A., "A Tandem Queue with Blocking and Markovian Arrival Process," *Queueing Systems*, vol. 41, no. 4, pp. 343–370, 2002.

[28] Green, J. V., "How many hospital beds?," *Inquiry*, vol. 39, pp. 400–412, 2002.

[29] Gupta, D. and Wang, L., "Revenue Management for a Primary-Care Clinic in the Presence of Patient Choice," *Operations Research*, vol. 56, pp. 576–592, May 2008.

[30] Gurvich, I., Armony, M., and Mandelbaum, A., "Service-Level Differentiation in Call Centers with Fully Flexible Servers," *Management Science*, vol. 54, pp. 279–294, Feb. 2008.

[31] Hagtvedt, R., *Applications of Decision Analysis to Health Care*. PhD thesis, Georgia Insitute of Technology, 2008.

[32] Hagtvedt, R., Ferguson, M., Griffin, P., Jones, G. T., and Keskinocak, P., "Cooperative Strategies to Reduce Ambulance Diversion," *Proceedings of the 2009 Winter Simulation Conference*, pp. 1861–1874, 2009.

[33] Haraden, C. and Resar, R., "Patient Flow in Hospitals: Understanding and Controlling It Better," *Frontiers of Health Services Management*, vol. 20, no. 4, pp. 3–15, 2004.

[34] Harrison, G. W., Shafer, A., and Mackay, M., "Modelling Variability in Hospital Bed Occupancy," *Health Care Management Science*, vol. 8, pp. 325–334, Nov. 2005.

[35] Harrison, J. M. and Lopez, M. J., "Heavy traffic resource pooling in parallel-server systems," *Queueing Systems*, vol. 33, pp. 339–368, 1999.

[36] Helm, J. E., Ahmadbeygi, S., and Van Oyen, M. P., "Design and Analysis of Hospital Admission Control for Operational Effectiveness," *Production and Operations Management*, vol. 20, no. 3, pp. 359–374, 2011.

[37] Hendrich, A., Fay, J., and Sorrells, A., "Effects of Acuity-Adaptable Rooms on Flow of Patients and Delivery of Care," *American Journal of Critical Care*, vol. 13, no. 1, pp. 35–45, 2004.

[38] Hodgins, M. J., Moore, N., and Legere, L., "Who Is Sleeping in Our Beds? Factors Predicting the ED Boarding of Admitted Patients for More Than 2 hours," *Journal of Emergency Nursing*, vol. 37, pp. 225–230, May 2011.

[39] Hoot, N. R. and Aronsky, D., "Systematic Review of Emergency Department Crowding: Causes, Effects, and Solutions," *Annals of Emergency Medicine*, vol. 52, pp. 126–136, Aug. 2008.

[40] Institute of Medicine, *Crossing the Quality Chasm: A New Health Systems*. Washington D.C: National Academy Press, 2001.

[41] Khuller, S., Mitchell, S. G., and Vazirani, V. V., "On-line Algorithms for Weighted Bipartite Matching and Stable Marriages," *Theoretical Computer Science*, vol. 127, pp. 255–267, May 1994.

[42] Kirby, A. and Kjesbo, A., "Tapping into Hidden Hospital Bed Capacity," *Healthcare Financial Management*, vol. 57, pp. 38–41, Nov. 2003.

[43] Lane, C. and Glassman, A., "Bigger And Better? Scaling Up And Innovation In Health Aid," *Health Affairs*, vol. 26, no. 4, pp. 935–948, 2007.

[44] Lasry, A., Zaric, G. S., and Carter, M. W., "Multi-level resource allocation for HIV prevention: A model for developing countries," *European Journal of Operational Research*, vol. 180, pp. 786–799, July 2007.

[45] Levin, S. R., Dittus, R., Aronsky, D., Weinger, M. B., Han, J., Boord, J., and France, D., "Optimizing cardiology capacity to reduce emergency department boarding: A systems engineering approach," *American Heart Journal*, vol. 156, pp. 1202–1209, Dec. 2008.

[46] Lippman, S. A., "Applying a New Device in the Optimization of Exponential Queuing Systems," *Operations Research*, vol. 23, no. 4, pp. 687–710, 1975.

[47] Mahdian, M. and Yan, Q., "Online Bipartite Matching with Random Arrivals: An Approach Based on Strongly Factor-Revealing LPs," in *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pp. 597–606, ACM, 2011.

[48] Mandelbaum, A., Momcilovi, P., and Tseytlin, Y., "On Fair Routing from Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers," *Management Science*, 2012.

[49] Mandelbaum, A., Momcilovi, P., and Tseytlin, Y., "On Fair Routing from Emergency Departments to Hospital Wards: QED Queues with Heterogeneous Servers," *Management Science*, 2012.

[50] Manshadi, V. H., Gharan, S. O., and Saberi, A., "Online Stochastic Matching: Online Actions Based on Offline Statistics," in *SODA 2011 Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1285–1294, 2011.

[51] Mayorga, M. E., Taaffe, K. M., and Arumugam, R., "Allocating Flexible Servers in Serial Systems with Switching Sosts," *Annals of Operations Research*, vol. 172, pp. 231–242, May 2009.

[52] McManus, M. L., Long, M. C., Cooper, A., and Litvak, E., "Queuing Theory Accurately Models the Need for Critical Care Resources," *Anesthesiology*, vol. 100, pp. 1271–1276, May 2004.

[53] McManus, M., Long, M., Cooper, A., Mandell, J., Berwick, D., Pagano, M., and Litvak, E., "Variability in Surgical Caseload and Access to Intensive Care Services," *Anesthesiology*, vol. 98, no. 6, pp. 1491–1496, 2003.

[54] Meyn, S. P., "Sequencing and Routing in Multiclass Queueing Networks Part I: Feedback Regulation," *SIAM Journal on Control and Optimization*, vol. 40, no. 3, p. 741, 2001.

[55] Meyn, S. P., "Sequencing and Routing in Multiclass Queueing Networks Part II: Workload Relaxations," *SIAM Journal on Control and Optimization*, vol. 42, no. 1, p. 178, 2003.

[56] Miller, B., "A Queueing Reward System with Several Customer Classes," *Management Science*, vol. 16, no. 3, pp. 234–245, 1969.

[57] Milne, E. and Whitty, P., "Calculation of Need for Paediatric Intensive Care Beds," *Arch Dis Child*, vol. 73, no. 6, pp. 505–507, 1995.

[58] Nooner, K., "Bouncing patients?," *Nursing*, vol. 35, no. 8, pp. 30–31, 2005.

[59] Ormeci, E. L. and Van Der Wal, J., "Admission Policies for a Two Class Loss System," *Stochastic Models*, vol. 17, pp. 513–540, 2001.

[60] Palmer, J. and Mitrani, I., "Optimal Server Allocation in Reconfigurable Clusters with Multiple Job Types," in *Computational Science and Its Applications ICCSA 2004*, no. December, pp. 76–86, Springer, 2004.

[61] Patel, P. B., Derlet, R. W., Vinson, D. R., Williams, M., and Wills, J., "Ambulance Diversion Reduction: The Sacramento Solution," *The American Journal of Emergency Medicine*, vol. 24, pp. 206–213, Mar. 2006.

[62] Pedroja, A. T., "The Tipping Point: The Relationship Between Volume and Patient Harm," *American Journal of Medical Quality*, vol. 23, no. 5, pp. 336–341, 2008.

[63] Perry, O. and Whitt, W., "Responding to Unexpected Overloads in Large-Scale Service Systems," *Management Science*, vol. 55, pp. 1353–1367, June 2009.

[64] Pham, J. C., Patel, R., Millin, M. G., Kirsch, T. D., and Chanmugam, A., "The Effects of Ambulance Diversion: A Comprehensive Review," *Academic Emergency Medicine*, vol. 13, pp. 1220–1227, Nov. 2006.

[65] Philip, A. J., *An Assessment of Equity on Geographical Allocation of Resources Relative to Need, in Public Primary Healthcare Services in the Northern Cape in South Africa*. PhD thesis, 2004.

[66] Proudlove, N. C., Black, S., and Fletcher, A., "OR and the Challenge to Improve the NHS: Modelling for Insight and Improvement in In-Patient Flows," *Journal of the Operational Research Society*, vol. 58, July 2006.

[67] Proudlove, N., Boaden, R., and Jorgensen, J., "Developing Bed Managers: The Why and the How," *Journal of Nursing Management*, vol. 15, pp. 34–42, Jan. 2007.

[68] Puterman, M. L., *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ: John Wiley & Sons, 1994.

[69] Quick, J. D., Management Sciences for Health, and Action Programme on Essential Drugs and Vaccines (WHO), *Managing Drug Supply. The Selection, Procurement, Distribution, and Use of Pharmaceuticals*. West Hartford, CT: Kumarian Press, second ed., 1997.

[70] Ramirez-Nafarrate, A., Fowler, J., and Wu, T., "Design of Centralized Ambulance Diversion Policies Using Simulation-Optimization," in *Proceedings of the 2011 Winter Simulation Conference*, pp. 1251–1262, 2011.

[71] Ramirez-Nafarrate, A., Fowler, J., and Wu, T., "Bi-criteria Analysis of Ambulance Diversion Policies," in *Proceedings of the 2010 Winter Simulation Conference*, pp. 2315–2326, 2010.

[72] Ross, S. M., *Stochastic Processes*. New York, New York, USA: John Wiley & Sons, second ed., 1996.

[73] Ross, S. M., *Introduction to Probability Models*. New York, New York, USA: Academic Press, eighth ed., 2003.

[74] Serfozo, R. F., "An Equivalence Between Continuous and Discrete Time Markov Decision Processes," *Operations Research*, vol. 27, no. 3, pp. 616–620, 1979.

[75] Serfozo, R. F., "Markovian Network Processes: Congestion-Dependent Routing and Processing," *Queueing Systems*, vol. 5, pp. 5–36, Nov. 1989.

[76] Spivey, M. Z. and Powell, W. B., "The Dynamic Assignment Problem," *Transportation Science*, vol. 38, no. 4, pp. 399–419, 2004.

[77] Stidham, S., "Analysis, Design, and Control of Queueing Systems," *Operations Research*, vol. 50, no. 1, pp. 197–216, 2002.

[78] Stock, R., *Africa South of the Sahara: A Geographical Interpretation*. New York, New York, USA: Guilford Press, second ed., 2004.

[79] Tanyan, B., "Riders for Health: Health Care Distribution Solutions in Sub-Saharan Africa," 2007.

[80] Teow, K. L., El-Darzi, E., Foo, C., Jin, X., and Sim, J., "Intelligent Analysis of Acute Bed Overflow in a Tertiary Hospital in Singapore," *Journal of Medical Systems*, Jan. 2011.

[81] Trzeciak, S. and Rivers, E. P., "Emergency Department Overcrowding in the United States: An Emerging Threat to Patient Safety and Public Health," *Emergency Medicine Journal*, vol. 20, pp. 402–405, 2003.

[82] Vee, E., Vassilvitskii, S., and Shanmugasundaram, J., "Optimal Online Assignment with Forecasts," in *Proceedings of the 11th ACM Conference on Electronic Commerce - EC '10*, (New York, New York, USA), ACM Press, 2010.

[83] Wilson, D. P. and Blower, S. M., "Designing Equitable Antiretroviral Allocation Strategies in Resource-Constrained Countries," *PLoS Medicine*, vol. 2, no. 2, p. e50, 2005.

[84] Wilson, D. and Kahn, J., "Predicting the Epidemiological Impact of Antiretroviral Allocation Strategies in KwaZulu-Natal : The Effect of the Urban Rural Divide," *Proceedings of the National Academy of Sciences of the United States*, vol. 103, no. 38, pp. 14228–14233, 2006.

[85] World Health Organization, "Health-for-All Policy for the 21st Century in the African Region: Agenda 2020," tech. rep., 2000.

[86] Xiong, W., Hupert, N., Hollingsworth, E., O'Brien, M., and Fast, J., "Can Modeling of HIV Treatment Processes Improve Outcomes? Capitalizing on Operations Research Approach to the Global Pandemic," *BMC Health Services Research*, vol. 8, 2008.