REDUCING HOSPITAL CONGESTION THROUGH IMPROVED INPATIENT DISCHARGE AND POST-ACUTE PLACEMENT: A STOCHASTIC PROGRAMMING APPROACH

A Dissertation

by

MARYAM KHATAMI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Mark Lawley
Committee Members,	Lewis Ntaimo
	Kiavash Kianfar
	Jon Stauffer
Head of Department,	Lewis Ntaimo

August 2020

Major Subject: Industrial Engineering

ABSTRACT

Hospital congestion is a pervasive problem that causes care delays, frustration for patients and their families, and stressed staff. This could potentially reduce the quality of care and ruin a hospital's reputation. Hospital congestion also affects new patient admission, mainly through the emergency department (ED). ED overcrowding, which has been a challenge for years, is primarily caused by patients waiting in ED for being admitted to the inpatient unit (IU). The main reason for the delay in ED patient transfer to IU is inpatient bed unavailability, which can also contribute to canceling elective surgeries and rejecting patient admission to intensive care units (ICU). The situation worsens with a pandemic virus outbreak, which boosts demand for ED and ICU beds. Thus, improving access to IU beds helps smooth the patient flow not only from the ED but also from other upstream units such as ICU and post-anesthesia care unit. One efficient way to release IU beds is to improve the discharge process and minimize non-medical inpatient days. This dissertation studies improving hospital discharge in both operational and strategic levels.

Discharge planning on the day of discharge is necessary to ensure effective performance. Discharge delay reduces patient satisfaction and increases hospital congestion and length of stay. Patient satisfaction is impacted by adherence to patient preferred discharge time. Preferences arise from many factors including waiting for family, avoiding rush hours, or waiting to feel better. Flow congestion manifests in patient boarding, and length of stay is extended if discharge delay incurs extra overnight stay. These factors are often in conflict, thus, good hospital performance can only be achieved through careful balancing. In the first part of this dissertation, discharge planning problem is formulated as a two-stage stochastic program with uncertainty in discharge processing and bed request times. The objective minimizes a combination of discharge lateness, patient boarding, and deviation from preferred discharge times. Patient boarding is integrated by aligning bed requests with bed releases. The model is solved for different instances generated using data from a large

hospital in Texas. Stochastic decomposition is compared with the deterministic equivalent and the L-shaped algorithm. A shortest expected processing time heuristic is also investigated. Computational experiments indicate that stochastic decomposition outperforms the L-shaped algorithm and the heuristic, with a significantly shorter computational time and small deviation from optimal. The L-shaped method solves only small problems within the allotted time budget. Simulation experiments demonstrate that the developed modeling approach improves discharge lateness and patient boarding compared to current practice.

In addition to patients being discharged to home, some wait for a transfer to the next level of care. These patients may experience several days of non-medical stay in IU until the hospital finds a post-acute care facility that fits their needs. The second part of this dissertation studies the feasibility of creating a "post-discharge-unit" (PDU) for patients, who are medically ready for discharge but are being delayed for some reason, to improve access to valuable IU beds. We use a multistage stochastic program to address PDU capacity planning and cost-effectiveness issues. The random variable is the number of bed requests from upstream units, including the ED, ICU, direct admissions, etc. Our model takes the impact of PDU on upstream patient flow, e.g., ED congestion and hospital admission into account. We use the stochastic dual dynamic programming algorithm to solve the model. An extensive numerical analysis is carried out using the data from a large hospital in Texas. An analysis of the impact of a variety of parameters, including PDU's fixed and operational costs, and length of alternate-level-of-care (ALC) stays, on PDU capacity and cost savings is performed. The results show that a PDU is cost-efficient and improves access to IU beds significantly, even when the ALC population is small, which is counter-intuitive. Another important finding is that PDU size in hospitals with a larger ALC population is more sensitive to increasing the PDU fixed and operational costs. In other words, the PDU size decreases faster when ALC population is larger.

DEDICATION

I dedicate this dissertation to the memory of my loving baby, Ali, whom I lost while working on this research, and to his little brother, Amirali, who gave a new meaning to my life. I will carry your love forever in my heart.

ACKNOWLEDGMENTS

I would like to thank my husband, Ashkan, for his incredible support during my difficult times. I would not have finished this without him. I would also like to thank my parents, Ommolbanin and Yousef, for their unconditional and endless support, love, and encouragement.

Thank you to Mohamed Ait, Davis Bivens, Jeremy Meade, and the nurses at MD Anderson Cancer Center in Houston, who helped me understand the discharge process and provided access to the related data.

Foremost, I would like to express my sincere gratitude to my advisor Dr. Mark Lawley for being an excellent advisor and for his continuous support, understanding, and encouragement during this process. I appreciate the countless hours he spent educating me and for always believing in me. I would like to thank my committee members Dr. Ntaimo, Dr. Kianfar, and Dr. Stauffer for their time and expertise. I am especially grateful to Dr. Ntaimo for his help and support during my Ph.D. studies. He taught me many things academically and personally. I also sincerely appreciate the help that I received from Dr. Stauffer, Dr. Kong, and Dr. Alvarado for completion of my research work.

NOMENCLATURE

ALC	Alternate level of care
AnS	Ancillary services
BaS	Basic services
BeC	Bed cleaning
DEP	Deterministic equivalent problem
EBR	Expected bed request
ED	Emergency department
EDP	Expected discharge processing
FiR	Final review
IDP	Inpatient discharge planning
ICU	Intensive care unit
IU	Inpatient unit
LTC	Long term care
MSP	Multistage stochastic program
PACF	Post-acute care facility
PACU	Post-anesthesia care unit
PDU	Post-discharge unit
RFD	Ready for discharge
SD	Stochastic decomposition
SDDP	Stochastic dual dynamic programming
SEPT	Shortest expected processing time
SPT	Shortest processing time

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Hospitals provide an array of services delivered by functional units such as emergency departments (ED), intensive care units (ICU), and inpatient units (IU). Lack of efficient coordination among these units results in hospital congestion which mainly affects patient admission through ED. Over the past decade, ED crowding has been a serious and pervasive problem. A 2002 hospital survey revealed that 90% of EDs in the U.S. operate at or over capacity [1]. This is due to falling ED capacity (4,960 in 1994 to 4,408 in 2014) and increasing ED visits (90M in 1994 to 136M in 2014) [2]. Consequences include critical care delays, patients leaving without care, and ambulance diversions to other EDs. Further, ED crowding negatively impacts a hospital's reputation and patients' satisfaction with the care process [3]. The main cause of ED crowding is ED boarding [4], the situation where patients unnecessarily occupy ED beds due to transfer delays in moving patients to inpatient units. In the U.S., ED patients wait roughly 3 hours for an available IU bed. This time increases to over 5 hours when EDs are overcrowded [1]. The main reason for ED boarding is IU bed unavailability [1], which can also contribute to canceling elective surgical operations [5] and rejecting patient admission to intensive care units [6]. Thus, any effort to release IU beds is extremely valuable.

An efficient approach to improve access to IU beds is the elimination of unnecessary IU stays, the primary source of which are patients waiting to be discharged. Some of these patients are ready for discharge (RFD) to home or the next readily available level of care. Improving the daily discharge process aimed at avoiding delays in operational activities will expedite IU bed release. Another part of unnecessary IU stays is related to patients waiting for transfer to post-acute care facilities (PACFs). They might experience several non-medical inpatient days due to the hospital's failure to find a PACF that fits the patient's needs. This unnecessary time in the hospital is referred to as the alternate level of care (ALC) days [7]. This dissertation studies the impact of both operational and strategic decision making on reducing non-medical IU stays of RFD and ALC patients.

1.1 Daily Inpatient Discharge Planning of RFD Patients

Patients often require services from several units of a hospital during an episode of care. Poor coordination between functional units is a major challenge leading to delays in care, patient flow congestion, and increased costs. Thus, coordination planning is essential for effective hospital performance. IU patients can be categorized into two classes, medical and surgical (based on discussions with nurses in several hospitals). Medical patients are admitted for illnesses that require nursing care. These patients exhibit a broad range of care needs and often require unexpected services prior to discharge. In contrast, surgical patients are usually in the IU to recover after a surgical procedure. Barring any surgical infection or complication, these patients are often discharged fairly quickly (2-3 days) with a rehabilitation regime that is typically managed at home. As a result, surgical patients are generally quicker to discharge with fewer unexpected complications.

Typically, the list of ready for discharge (RFD) patients is reasonably known one day ahead, but the discharge *time-of-day* is uncertain due to staff availability, emerging patient need, variable discharge processing times, transportation arrangement, and so forth. Discharge staff must consider and account for this wide variety of circumstances. Further, patients often prefer either early or late discharge times for a variety of reasons. For example, patients may prefer to leave certain times during the day to avoid heavy traffic or later if they feel anxious about their health. Meeting these preferences helps improve satisfaction with the overall hospital stay. Unfortunately, little research exists to support decision making for discharge processes.

This research develops a stochastic parallel-machine sequencing-scheduling model for the inpatient discharge planning (IDP) problem. Patients are modeled as jobs while nurses or other resources are treated as machines. This approach is apt because nurses are typically independent and work in parallel. The IDP problem is formulated as a two-stage stochastic mixed-integer program with two sources of randomness; uncertain discharge processing times and uncertain bed request arrival times. *The method assigns every patient to a nurse,*

determines the optimal sequence of patients, and assigns IU beds to bed requests such that the total penalty of violating patient preference, discharge lateness, and patient boarding are minimized. Given the two-stage structure, the L-shaped algorithm and stochastic decomposition (SD) are applicable solution methods. These algorithms are used to solve instances of the IDP problem generated using real data from a large hospital in Texas. A shortest expected processing time heuristic (SEPT), taken from the machine scheduling literature, is also tested on the IDP problem. Contributions of the work include: (1) mathematical formulation of the IDP problem optimizing adherence to patient preferences and hospital performance, with uncertainty in discharge processing time and ED bed request arrivals; (2) solution algorithm implementation including the L-shaped algorithm, a sophisticated interior sampling technique (SD), and a modified SEPT heuristic; (3) numerical analysis, using real world data, comparing algorithmic performance metrics including solution quality, computational speed, and sensitivity analysis.

1.2 Managing Non-medical Inpatient Stays of ALC Patients

One efficient strategy to improve access to IU beds is reducing unnecessary inpatient hospital stays. An example is IU bed occupancy by patients who are medically ready for discharge but waiting for transfer to a lower level of care (ALC patients). While the majority of inpatients are discharged to their homes, some are waiting on transfer to post acute care facilities (PACFs) including rehabilitation centers, long-term care hospitals, and skilled nursing facilities. Patients recovering from a stroke, hip fracture, and neurological diseases typically need post-acute care. More than one-third of stroke patients in the US are discharged to PACFs [8]. Hospitals and PACFs failing to efficiently coordinate is the main reason for delayed discharges [9, 10], leading to increased ALC days. [11] showed that 31.8% of long-stay inpatient days were non-medical in The Mount Sinai Hospital in New York. This was mainly caused by transfer delays to nursing facilities.

Hospitals face several challenges when transferring complicated patients to PACFs, typically due to lack of facilities that fit each patient's needs. [12] showed that patients spend

12% of their hospital stays waiting for a rehabilitation bed. Another study [13], based on data from a large Canadian hospital, found that 41.5% of unnecessary hospital stays are related to patients waiting for admission to skilled nursing facilities, while these patients only account for 8.8% of ALC patients. Based on our interviews with leadership at a large cancer hospital in Houston, TX, PACFs frequently refuse to admit complicated cancer patients since they typically need specific and expensive medications and physicians with particular specialties. Even when a suitable PACF is found, miscommunication between the hospital and the PACF or busyness of the PACF staff might delay the patient's transfer. In such cases, patients have to stay in the IU while medically ready for discharge, which contributes to congestion in other units such as the ED, ICU, and PACU. [8] found that the average waiting time for patient replacement in an academic inpatient neurology ward was 4.8 days, per ALC patient, including a worst-case of 80 days for one patient. Thus, improving patient transmission to PACFs is essential to release IU beds for upcoming bed requests. A bed request is made when a patient from an upstream unit (e.g., ED, ICU, and PACU) or from outside the hospital (direct admission) needs an inpatient bed and is ready for being transferred to IU.

Opening a post-discharge unit (PDU) for discharged patients, waiting on transfer to PACFs, can avoid unnecessary occupancy of valuable IU beds. The PDU not only improves patient flow but also decreases hospital costs in many ways by avoiding needless services. First, utilizing a PDU reduces costs by increasing patient to nurse ratios and reducing physician coverage. Second, patients in the PDU do not require as much medical or diagnostic equipment in the room. Third, they typically need more non-hospital services which can be delivered by case managers, social workers, and physical therapists. In fact, physical therapy is the primary need for patients who require post-acute care after discharge. Thus, centralizing these patients in a PDU improves the rounding process of physical therapists, case managers, and social workers visiting every IU in the hospital.

From the patients' perspective, discharge delays are frustrating and uncertain [14]. Based

on patient experience studies [15], ALC patients feel isolated and neglected, and the nursing staff's lack of attention ruins their hospital-stay experience. Transferring ALC patients to the PDU helps reduce the feeling of being ignored. Some patients travel long distances to be treated in a specific hospital or by a particular physician. Thus, they may not feel ready for transfer to a different facility or may prefer to stay closer to the hospital in case their condition worsens. Further, patients and their family caregivers need to be educated on what to expect at PACFs. However, this often does not happen due to time limitations. In fact, most patients feel rushed when transitioning care [16]. Inappropriate preparation of patients could confuse them and result in failure to manage their recovery treatment, which could ultimately lead to a readmission. [17] showed that encouraging older patients to participate in their care transition actively reduces re-hospitalization. Staying in the PDU gives patients, their family caregivers, and the nursing staff more time to prepare patients for this transition. Last but not least, patients could benefit from this step down in care to the PDU financially. According to our healthcare professional partners, the PDU chargeable rates should be lower than the IU rates since the level of care is different.

Considering the aforementioned benefits of a PDU, this study aims to understand how opening a PDU could smooth patient flow and improve hospital operations. Therefore, several research questions are proposed to investigate this. *First, what is the optimal capacity of the PDU, and how much would it cost? Second, how much does an optimally sized PDU improve access to IU beds? Third, what is the magnitude of overall cost saving to the hospital?* This research is the first in the operations research literature to study patient transfer from hospitals to post-acute care facilities and its impact on patient flow in the hospital. A multistage stochastic program (MSP) is developed for capacity planning and cost-effectiveness analysis of a post-discharge unit. The PDU's optimal capacity is impacted by the number of bed requests from upstream units, which is a random variable due to the uncertainty in patient arrivals to the ED, ICU, and PACU. Stages in the MSP are defined as t-duration time intervals where t is based on hours. The first-stage decision determines the capacity of

the PDU, while the next stage decisions are operational decisions based on the number of patient admissions to the IU and patient transfers to the PDU. The stochastic dual dynamic programming (SDDP) algorithm is adapted to solve the model using actual data from a large hospital in Texas.

Contributions of the work include: (1) addressing the issue of ALC patients' non-medical stays for the first time in the operations research literature; (2) mathematical formulation of the problem optimizing hospital cost and performance, with uncertainty in number of bed requests from upstream unit; (3) numerical analysis, using real world data, investigating the optimal PDU size, improvement in IU access, and cost savings.

The rest of this dissertation is structured as follows: Chapter 2 reviews the literature of hospital discharge, patient transition to post-acute care facilities, and related stochastic programming methods. Chapter 3 presents the IDP problem setting, the optimization framework, the solution algorithm, and numerical results. The MSP mathematical model for capacity planning and cost-effectiveness of a PDU unit is developed, verified, and analyzed in Chapter 4. Chapter 5 concludes the dissertation and proposes future research directions.

2. LITERATURE REVIEW

Lack of smooth patient flow is the main contributor to hospital congestion. Patient flow is widely studied in the operations research (OR) community. In general, the literature of patient flow includes three main areas of inflow/admission, inside hospital operations, and outflow/discharge. There is a vast literature on the first two. A few examples include outpatient scheduling [18, 19, 20], ED overcrowding management [21, 22, 23], hospital admission scheduling and strategies [5, 24], and surgery scheduling [25, 26]. However, the role of hospital outflow/discharge on patient flow and hospital congestion is almost missing in the OR literature. This chapter reviews patient flow modeling literature with a focus on patient outflow/discharge.

2.1 Operational Decision Making: Improvement of Inpatient Discharge Process

Several authors investigate the impact of discharge delay on IU bed availability, hospital costs, and patient length of stay. [27] and [28] state that IU beds are often unavailable due to discharge delays caused by inefficient IDP. [29] found that 11.9% to 36.7% of hospital stays experience discharge delays, raising hospital costs by 9%. This increase results from extra overnight stays, ED congestion and ambulance diversion, and overworked staff. A study at a 233-bed tertiary-care children's hospital indicated that the system-wide effect of poor IDP was 82 delay-related inpatient days (9% of total) and $170,000 (8.9%) in excess costs over a one-month period [30]. The study found that 25% of patients experienced at least one day of delay, with 58% of delays caused by incomplete discharge tasks or lack of follow-up by ancillary services (such as physical, speech, and radiation therapy). Several researchers contend that inpatient discharge delays are critical factors in ED overcrowding and boarding [27, 28]. [31] found that increasing inpatient beds reduces ED crowding more than increasing ED beds.

In the emergency medicine and health services literature, several strategies are proposed,

e.g., increasing discharge before noon percentage [32, 33], in-room display of discharge appointment [34], and admission-discharge balancing [28, 35]. However, none of these are rigorously modeled.

A few research papers in the operations research community focus on the importance of discharge strategies in hospital units. [36] develop a Markov chain model to reduce ICU overcrowding by discharging patients with the smallest remaining length of stay. The authors also investigate the impact of elective surgery schedules on ICU performance. [37] investigate the impact of ICU discharge strategies on readmission rate and mortality. They use dynamic programming to find an index policy that is proven to be optimal under some conditions and near-optimal otherwise. The policy assists ICU physicians in deciding on discharging a relatively stable patient to be able to admit a new one. [38] propose a stochastic network model for inpatient flow management with the purpose of minimizing ED boarding. Via simulation studies, the authors use their model to evaluate the impact of operational policies on ED boarding. The policies are increasing bed capacity, having a limit on maximum length of stay, and reducing allocation delays caused by reasons other than bed unavailability.

Some papers in the literature study the impact of non-discharge related strategies on reducing ED boarding. These strategies include ambulance diversion [39], ambulance redeployment and dispatching [40], controlling inpatient admissions [41, 42], ED resource capacity management [43], ED bed capacity management and ambulance diversions [44], ED bed capacity management and leave-without-treatment [45], streaming patients through ED based on the possibility of being discharged or admitted to the hospital [46], prioritizing patients based on their complexity [47], and statistical monitoring of the daily operation of EDs [23].

Although the research discussed above is related to this dissertation, none of the literature cited addresses discharge planning at the level of granularity considered in this study. Simple heuristics, such as increasing discharge before noon, do not capture the inherent stochastic structure that characterizes uncertainty. Further, the OR work that acknowledges uncertainty does not optimize patient flow across departments. In contrast, this disserta-

tion uses knowledge of uncertainty coupled with optimization approaches to improve system performance across hospital departments. Thus, it is relevant for researchers attempting to optimize patient flow within any specific department, e.g., ED.

2.2 Strategic Decision Making: Management of ALC Patients' Discharges

As explained in Chapter 1, non-medical inpatient stays (ALC days) are among primary reasons for patient congestion and inefficient patient flow. Some articles in the emergency medicine literature study characteristics of ALC patients and risk factors, but do not suggest a policy to resolve the issue of ALC days. The main focus of OR literature has been on capacity expansion of post-acute care facilities, particularly LTCs, as a policy to reduce hospital congestion caused by patients waiting for a transfer to PACFs. A few papers in operations literature analyze the impact of different policies on ALC population in hospitals. The following is a review of related papers in both emergency medicine and operations literature.

- *The Emergency Medicine and Health Services Literature:* Many articles study contributing factors to the prolonged length of stay of ALC patients, characteristics of these patients, and discharge barriers. Risk factors for designating as ALC include functional impairment and being medically complicated [48], older ages [7, 49], behavioral and financial barriers (e.g., being homeless) [50], co-morbidity [51], and inefficient communication with the next level of care [9, 49]. These papers, although valuable, do not suggest/model strategies to resolve the issue of ALC days and do not investigate the impact of these non-medical stays on patient flow in hospitals.

- *OR Literature on Capacity Planning of Post-acute Care Facilities:* There is broad literature in community care capacity planning, mainly focused on long term care (LTC) facilities, to reduce congestion in different facilities in healthcare systems [52, 53, 54, 55]. Although improving access to post-acute care facilities such as LTC could reduce ALC days in the hospital, the following shortcomings are worth attention. First,

these strategic changes are not in hospital's control, so the hospital does not have any authority to ensure such changes be implemented. Second, LTCs are only one type of post-acute care facilities. ALC patients are transferred to a variety of facilities such as skilled nursing facilities which contribute to ALC days the most. Third, although waiting for transfer to the next level of care is the main reason for non-medical inpatient days, it is not the only reason. Some patients stay for a long time in the hospital, after they are medically ready for discharge, due to financial and/or behavioral problems or lack of decision-making capacity [50]. Thus, these ALC days need to be managed to avoid unnecessary occupancy of IU beds as much as possible.

- *OR Literature on ALC Patients:* The following reviews a few available papers in the operations literature that investigate patient congestion in healthcare systems, focusing on reducing ALC days. [56] presents a queuing model to analyze the impact of discharge rates on the number of ALC days. The authors compare the performance of their model with the current practice in seven hospitals in New York. [57] develop a Markov decision process model to ensure the number of ALC patients remains under a threshold specified by the hospital. The ALC patients in their focused hospitals are mainly waiting for transfer to an LTC facility, which also admits patients from the community. The paper presents a policy which prioritizes hospital clients over community clients, as long as the census of ALC patients is above the target unless there is a an urgent community client. If the ALC population in the hospital is below the target, an available LTC bed is assigned to community clients with the longest waiting times. Although this policy meets threshold requirements for the ALC patient population, it limits the LTC's ability to respond to the community demand. Authors suggest reducing LTC length of stay as a strategy to cope with this consequence of the policy. They also develop a simulation model to compare their policy with other policies. [58], an extension to [57], develops a simulation model to investigate the possibility of meeting both the hospital's threshold for number of ALC patients and the target waiting

time of direct admissions from the community. The authors conclude that the available capacity in their studied region is not sufficient to achieve these goals. However, they show that a combination of increasing supportive housings (facilities that provide housing and medical services for people without stable housing and suffering from disabilities and diseases such as mental illness, HIV, etc.) and reducing patients' length of stay in LTCs (from an average of 3 years to 2 years) meets both targets. [59] use queuing network models with blocking to study the impact of different facility capacities on their congestion in the Philadelphia mental health system. The facilities they investigate include acute care settings, extended acute hospitals, residential facilities, and supported housing. The main finding of the study is that the Philadelphia health system needs to increase the capacity of supported housing to prevent the tremendous blockage of its mental facilities in the future.

2.3 Relevant Stochastic Programming Methods

- *Two-stage Stochastic Programming for Parallel Machine Scheduling:* The IDP problem is similar to parallel machine scheduling with stochastic job processing times. Several problems with similar structures are studied in the healthcare literature. [60] formulate the multi-server appointment scheduling problem with uncertain service durations as a two-stage mixed-integer program with chance constraints limiting the risk of server overtime. The authors develop a decomposition algorithm to solve this problem. [61] formulate the ED environment as a two-layer model. The first layer assigns patients to ED beds using a parallel machine scheduling framework. In the second layer, a flexible job shop model is developed to assign tasks for assisting ED patients to resources. A heuristic method is proposed to solve the model. [62] study the surgery to operating room allocation with uncertain surgery duration. They formulate this problem as a two-stage stochastic program and apply an adapted integer L-shaped solution method. [63] model the operating room scheduling problem with uncertain surgery duration as a chance-constrained stochastic program. Joint chance constraints guarantee admissible

levels of operating room overtime.

In all of these cases, scheduled services are similar to jobs with uncertain service durations and multiple service resources are similar to machines working in parallel. The IDP problem also exhibits these similarities with a patient discharge being equivalent to a job with uncertain processing time and a nurse being equivalent to a machine. In addition, patient discharges have uncertain due dates represented by the random time at which the occupied bed will be needed (the time at which bed requests occur).

- *Multistage Stochastic Programming*: This dissertation is the first one that applies optimization tools for providing hospitals with implementable policies to manage ALC patients, improve patient flow, and reduce costs. Nested multistage stochastic programs (MSPs) fit the best with this problem as they capture the uncertainty and cope with outcomes evolving sequentially over time. Multistage stochastic programs (MSPs) are widely used in supply chain related problems such as production planning, inventory control, lot sizing, and capacity planning [64, 65, 66]. MSPs are also popular in energy management and power generation [67, 68]. Finance problems, more specifically asset allocation, are another well-known application for MSPs [69]. The literature of MSP applied to healthcare, however, is narrow. Examples include infectious disease control [70], clinical trial of new drugs [71], nursing staff allocation [72], and appointment scheduling [73].

3. INPATIENT DISCHARGE PLANNING

3.1 Problem Description

3.1.1 Discharge Process

Generally, the list of RFD patients is known at the beginning of each day. To be discharged, these patients need several services, which are categorized as basic (BaS), ancillary (AnS), and final review (FiR). For example, required basic services might include self care education and medication reconciliation, whereas required ancillary services might include therapy sessions or diagnostic tests. While surgical patients often need only some basic services, more complex medical patients typically require additional ancillary services and thorough final review by the attending physician. After discharge, the patient's bed is released and must be cleaned (BeC) before being assigned to a bed request from upstream units. Figure 3.1 provides an overview of discharge process.

To avoid unnecessary overnight stays, hospitals typically prefer to discharge patients by a specific time of day, referred to as the *target discharge time*. Although only a soft deadline, the target time helps in managing patient flow. However, predicting discharge time is often difficult, especially for medical patients, since discharge processing time varies by patient complexity, availability of hospital resources, and patient transportation.

Patient satisfaction is a widely held indicator of the quality of hospital service. It enhances patient retention and loyalty and improves hospital reputation and profit. Research has shown that the discharge process has a significant impact on patient satisfaction, with discharge time of day preferences being a key factor [74]. These preferences arise for many reasons. For example, patients may not feel ready for discharge. They may still feel sick or incapable of self care. Such patients prefer the latest possible discharge time. Other patients may prefer to leave before or after traffic hours, or may need to accommodate family member schedules. Thus, it is important to focus not only on the hospital's target discharge time

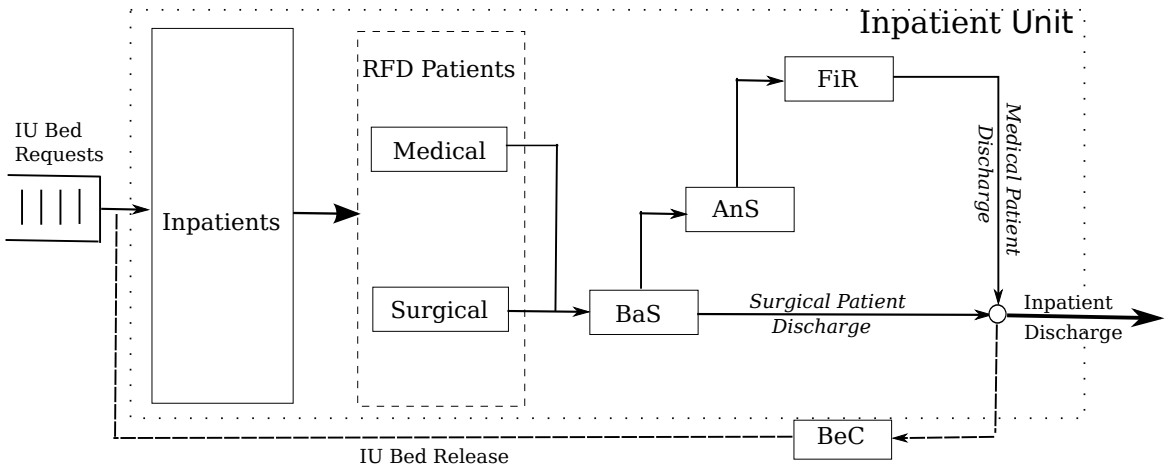


Figure 3.1: Overview of Discharge Process

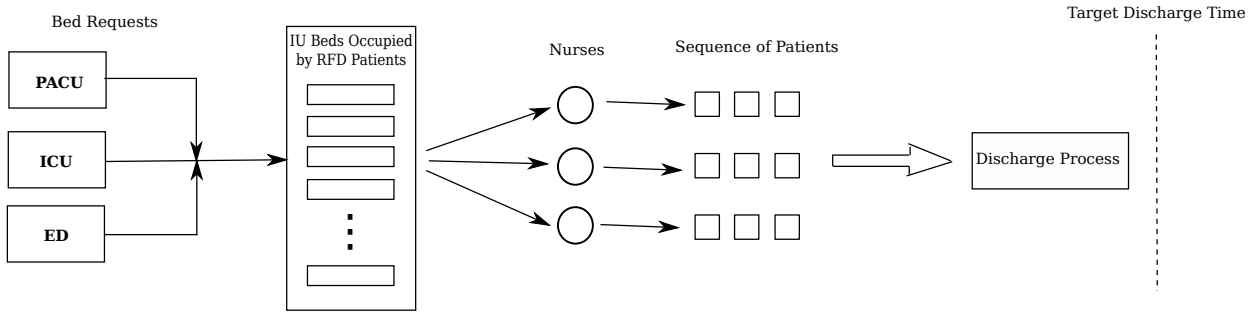


Figure 3.2: Structure of the IDP Problem

but also on accommodating patient preferences on discharge time of day.

3.1.2 IDP Modeling Assumptions

This section develops the underlying modeling assumptions for optimally assigning RFD patients to nurses and sequence their discharges. In addition, the model will assign an IU bed to each bed request from other units. The model minimizes total expected discharge lateness (positive deviation of a patient's discharge time from hospital's target time) and patient boarding (positive deviation of IU bed availability time from bed request time) in upstream units. Figure 3.2 provides a schematic of the IDP problem. The IDP model makes the following assumptions.

1. RFD patients are known at the beginning of the day.
2. Patients may have preferences on discharge time of day.
3. Patient to nurse assignments are constant throughout a shift.
4. Number of IU nurses is known, nurses are identical, even across shifts.
5. Discharge process time and bed request arrival time are independent random variables.
6. Bed requests up to number of discharges are considered.
7. Stochastic discharge processing times are exogenous from the first-stage decisions.
8. Steps BaS, AnS, FiR, and BeC are aggregated into one step.

Assumptions 1-4 are consistent with the IU environment. If discharge processing times exceed shift length, then a new, equally skilled, nurse assumes the responsibilities of the outgoing nurse. Any delay incurred during the shift change is captured in the historical data for stochastic discharge processing times. Assumption 5 is a basic requirement of stochastic programming and is reasonable in the IDP setting.

Assumption 6 implies that the maximum number of bed requests that can be satisfied is equal to the number of RFD patients. In practice, excess bed requests will be diverted to other units and do not impact the discharge process for the current unit. Since our model is not designed for capacity planning purposes (e.g. making decisions on the number of beds in IU), assumption 6 does not make a difference in the decision-making process.

Assumption 8, aggregation of steps, requires a more detailed justification.

- *Inconsistent AnS step*: Not all patients require the same services during the AnS step and often the exact AnS needs are not known at the beginning of the day.
- *External resources*: Even if regular ancillary services are known, the services are external to the decision-making unit and thus there is little to no control over their availability.

- *Overlapping steps*: The steps of the discharge process overlap. For example, a physician can review results of BaS while the patient is in the AnS department.
- *Nurse multi-tasking*: The IDP model assumes sequential patient discharges. In reality, when a patient is sent to the AnS department, the nurse can start BaS for next patient.
- *Emphasis on final completion time*: The objective function (discharge lateness and patient boarding) does not require the detailed individual starting and stopping time of every step, but rather the completion time of the final step (discharge completion time) for each patient.
- *Historical data*: The historical data for discharge processing times captures the complexity of the previously mentioned confounding issues.
- *Gamma distribution*: The processing times of all steps (BaS, AnS, FiR, BeC) are independent, random, and constitute the total discharge processing time. Based on best fit computations using historical data, the total discharge processing time is well represented by a gamma distribution, which is commonly used to represent random service times. Since a service is typically composed of a variety of constituent activities determined by customer need (imagine a teller who cashes a check for one customer and takes a deposit from the next), it is clear that the gamma of the total service time often provides adequate modeling representation of the composition of its random constituents. This work makes this common assumption.
- *Practical impact*: To test this assumption, we used a simulator (discussed in Section 3.4.3) to experiment with aggregation of the four-step IDP. In testing, we found little practical difference in the performance metrics between simulation with four steps and simulation with one (Tardiness 1.2%, Boarding 0.1%, Penalty 0.8%). Further, note that the second-stage model, (3.2a)-(3.2f), is a linear program (LP) that captures first-stage solution performance under given scenarios. All constraints in that LP are intrinsically enforced in the operation of the simulator, while the simulator captures additional

detail that the LP does not. Thus, it is reasonable to assume that if the more detailed simulator indicates little difference in performance under aggregation, then the LP will likely reflect the same. This indicates that for the IDP problem, second-stage evaluation errors due to aggregation are not likely to be of practical importance, especially given the uncertainties discussed above.

3.2 Two-Stage Stochastic Program

The IDP problem is formulated as a two-stage stochastic mixed-integer program. Stochastic programs are useful when one or more aspects of the constraints or the objective function is uncertain. Problem (P) presents the general structure of the two-stage stochastic program with recourse where the uncertainty is represented by $\tilde{\omega}$.

$$\begin{aligned} (P) \qquad \text{Min} \quad & c^\top x + \mathbb{E}[f(x, \tilde{\omega})] \\ \text{s.t.} \ & x \in X, \end{aligned}$$

where for every outcome ω of random variable $\tilde{\omega}$, $f(x, \omega)$ is defined in the following:

$$\begin{aligned} f(x, \omega) = \text{Min} \quad & q(\omega)^\top y(\omega) \\ \text{s.t.} \ & W y(\omega) \ \geq r(\omega) - T(\omega)x \\ & y(\omega) \geq 0. \end{aligned}$$

First stage decisions, represented by the $n_1 \times 1$ vector x, are made before realization of the uncertainty. Note that c is the cost coefficient vector of size $n_1 \times 1$. In the second stage, recourse actions represented by the $n_2 \times 1$ vector $y(\omega)$ are taken for a given realization ω. The second-stage data include $q(\omega)$, $r(\omega)$, and $T(\omega)$ which are respectively of sizes $n_2 \times 1$, $m_2 \times 1$, and $m_2 \times n_1$. Each of this data could be random. The $m_2 \times n_2$ matrix W is called the recourse matrix and is fixed. Figure 3.3 visualizes the the decision-making process using two-stage stochastic programming.

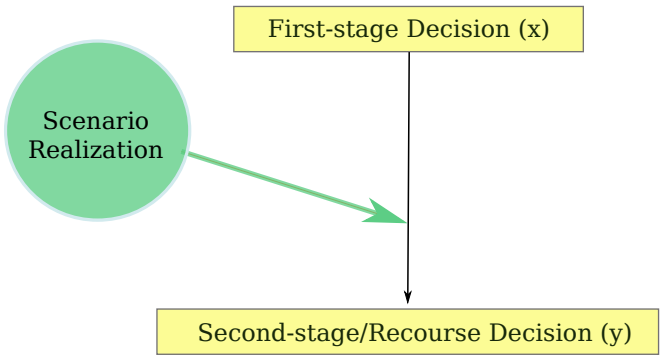


Figure 3.3: Two-stage Stochastic Programming

The IDP first-stage model sequences and assigns RFD patients to nurses and assigns IU beds to bed requests. Recourse decisions are discharge completion times for each RFD patient and bed availability times. As explained above, the second-stage model evaluates the

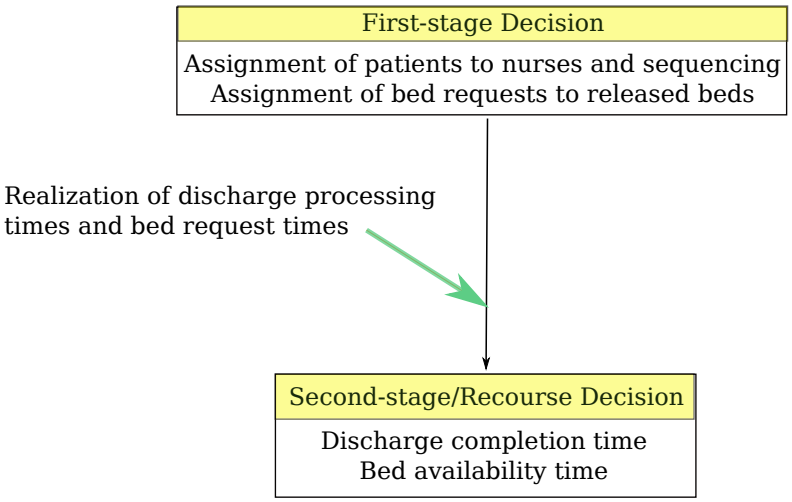


Figure 3.4: First and Recourse Decisions in IDP Model

performance of the first-stage decisions for each scenario. In the IDP problem, each scenario

consists of a realization of discharge processing time for each RFD patient and the arrival time of each bed request. These are represented in Figure 3.4.

3.2.1 Notation

This section defines the IDP model notation and presents the mathematical formulation. Table 3.1 presents set notation, most of which is obvious. The set, K, bears some explanation. K is the set of positions in the sequence of patients assigned to every nurse, (in Figure 3.2, each has three positions, thus $K = \{1, 2, 3\}$). The size of set K, $|K|$, specifies the maximum number of patients that a nurse can discharge. Positions of the patients in the sequence determine the order in which the nurse will discharge them. Table 3.2 presents the first-

Table 3.1: Sets for the IDP Model

I	Set of nurses, indexed by i
J	Set of RFD patients, indexed by j
M	Set of bed requests, indexed by m
K	Set of sequence positions, indexed by k
Ω	Set of scenarios, indexed by ω

stage parameters and decision variables. The first-stage model minimizes total penalty of violating patient preferences for discharge time. For each patient, a penalty is assigned to every position in the sequence with respect to preferred discharge time. For example, if a patient prefers to be discharged early in the morning, the penalty of assigning the patient to the first position is 0. There is a penalty for assigning the patient to any other position. The larger the deviation of the position from the patient's preferred one, the greater the penalty. The parameter η determines the importance of meeting patient preferences versus minimizing discharge lateness and patient boarding. Satisfying patient preferences may be as important as minimizing discharge lateness and patient boarding ($\eta = 1$), less important ($\eta \in (0, 1)$), or not important ($\eta = 0$).

Table 3.2: First-Stage Parameters and Decision Variables

First-Stage Parameters	
$\alpha_j^k \in [0, \infty)$	Penalty of assigning patient j to position k with respect to patient's preference
$\eta \in [0, 1]$	Importance of patients preferences compared to discharge lateness and patient boarding
First-Stage Decision Variables	
x_{ij}^k	1, if patient j is scheduled as the k^{th} patient for nurse i, 0 otherwise
u_{im}^k	1, if bed request m is assigned to the bed occupied by k^{th} patient for nurse i, 0 otherwise

Table 3.3 presents second-stage parameters and decision variables. These depend on first-stage decisions and realization of random variables. Discharge completion and bed availability times depend on assignment and sequencing decisions from the first stage and realization of discharge processing times. Bed request arrival times influence boarding time in upstream units.

3.2.2 Formulation

This section presents a two-stage stochastic program for the IDP problem. Equations (3.1a)-(3.1h) represent the first-stage model, and equations (3.2a)-(3.2f) represent the second stage. The first term in (3.1a) is the total penalty of deviating from patient preferred discharge times. The second term is the expected future cost (expected total discharge lateness and patient boarding). First-stage decisions are represented by the $(ijk) \times 1$ vector $X = (x_{ij}^k : \forall i, j, k)$ and the $(imk) \times 1$ vector $U = (u_{im}^k : \forall i, m, k)$. The function $f(X, U, \omega)$ is called recourse function which depends on first-stage decisions X and U, and a realization (scenario) ω of random variable $\tilde{\omega}$.

Equation (3.1b) ensures that every patient is assigned to only one nurse and one position in the sequence of patients assigned to that nurse. Constraint (3.1c) guarantees that at most

Table 3.3: Second-Stage Parameters and Decision Variables

Second-Stage Parameters			
$\tilde{\omega}$	Multivariate random variable	d	Target discharge time for the inpatient unit
ω	Realization (scenario) of $\tilde{\omega}$	$p_j(\omega)$	j^{th} patient's discharge processing time under ω
$r_m(\omega)$	m^{th} bed request time under ω	$\mathcal{M}$	A large number
w_m	Importance ratio of request m		

Second-Stage Decision Variables	
$c_{ij}^k(\omega)$	Discharge time of patient j if scheduled to k^{th} position for nurse i, under ω; 0 otherwise
$v_{im}^k(\omega)$	Bed release time of patient in position k of nurse i, under ω, if assigned to m^{th} request; 0 otherwise

one patient is assigned to each nurse and position pair. Constraint (3.1d) ensures that a position in the sequence can be filled only if the preceding position is filled. Equation (3.1e) enforces that every bed request is assigned to only one bed (the bed occupied by the patient assigned to nurse i, position k).

Constraint (3.1f) ensures that no more than one bed request can be assigned to every bed. Constraint (3.1g) requires that if no RFD patient is assigned to a nurse and position pair, there is no bed request assigned to that position (no patient is assigned that position, and thus the position releases no bed). Constraint (3.1h) enforces binary values for first-stage decision variables.

The first term in (3.2a) determines the total discharge lateness over all patients. The second term represents the total weighted lateness in providing empty beds for bed requests. Constraint (3.2b) ensures that $c_{ij}^k(\omega)$ can take a non-zero value only if patient j is scheduled as the k^{th} patient for nurse i. Constraint (3.2c) enforces that if $x_{ij}^k = 1$ then the discharge

process for patient j starts after all patients sequenced earlier are discharged.

$$\text{Min} \sum_i \sum_j \sum_k \eta \alpha_j^k x_{ij}^k + \mathbb{E}(f(X, U, \tilde{\omega})) \tag{3.1a}$$

s.t.

$$\sum_k \sum_i x_{ij}^k = 1, \qquad \forall j \tag{3.1b}$$

$$\sum_j x_{ij}^k \leq 1, \qquad \forall i, k \tag{3.1c}$$

$$\sum_j x_{ij}^k \leq \sum_j x_{ij}^{k-1}, \qquad \forall i, k \geq 2 \tag{3.1d}$$

$$\sum_k \sum_i u_{im}^k = 1, \qquad \forall m \tag{3.1e}$$

$$\sum_m u_{im}^k \leq 1, \qquad \forall i, k \tag{3.1f}$$

$$u_{im}^k \leq \sum_j x_{ij}^k, \qquad \forall i, k, m \tag{3.1g}$$

$$x_{ij}^k, u_{im}^k \in \{0, 1\}, \qquad \forall i, j, k, m \tag{3.1h}$$

where for each realization $\omega \in \Omega$ of $\tilde{\omega}$,

$$f(X, U, \omega) = \text{Min} \sum_i \sum_j \sum_k (c_{ij}^k(\omega) - d)^+ + \sum_m \sum_i \sum_k w_m (v_{im}^k(\omega) - r_m(\omega))^+ \tag{3.2a}$$

s.t.

$$c_{ij}^k(\omega) \leq \mathcal{M} x_{ij}^k, \qquad \forall i, j, k \tag{3.2b}$$

$$c_{ij}^k(\omega) \geq \sum_{k'=1}^{k} \sum_{j'=1}^{|J|} p_{j'}(\omega) x_{ij'}^{k'} - \mathcal{M}(1 - x_{ij}^k), \qquad \forall i, j, k \tag{3.2c}$$

$$v_{im}^k(\omega) \leq \mathcal{M} u_{im}^k, \qquad \forall i, m, k \tag{3.2d}$$

$$v_{im}^k(\omega) + \mathcal{M}(1 - u_{im}^k) \geq \sum_j c_{ij}^k(\omega), \qquad \forall i, m, k \tag{3.2e}$$

$$c_{ij}^k(\omega), v_{im}^k(\omega) \geq 0, \qquad \forall i, j, k, m \tag{3.2f}$$

Constraint (3.2d) ensures that $v_{im}^k(\omega)$ can take a non-zero value only if the m^{th} bed request is assigned to the bed occupied by nurse i's k^{th} patient (in which case, $u_{im}^k = 1$), otherwise it is 0. Constraint (3.2e) guarantees that if $u_{im}^k = 1$, the bed will be free as soon as the patient in the bed is discharged. Constraint (3.2f) imposes non-negativity restrictions on second-stage decision variables.

3.2.3 An Upper Bound for Big-$\mathcal{M}$

The existence of big-$\mathcal{M}$ in the second-stage model (constraints 2b-2e) is not computationally desirable. This section tightens the formulation of the second-stage by finding an upper bound for the big-$\mathcal{M}$. The following facts are used to this end:

- The second-stage model is defined for every scenario ω and does not include a scenario-linking constraint. Therefore, an upper bound for the big-$\mathcal{M}$ is scenario-specific.
- An upper bound for the big-$\mathcal{M}$ in constraint (3.2b) is equivalent to an upper bound for $c_{ij}^k(\omega)$.
- The IDP problem is related to identical parallel machine scheduling. The processing time of each job is thus the same on every machine. This implies that the upper bound of the job completion time (which is equivalent to that of the big-$\mathcal{M}$) does not depend on the machine and only depends on the position of the job in the sequence.

Let $\mathbb{P}(\omega)$ be the vector of patient discharge processing times given scenario ω, $\mathbb{P}(\omega) = (p_j(\omega) : \forall j \in J)$. Define $\xi(\omega)$ as the following non-increasing sequence:

$$\xi(\omega) = \{\zeta_\ell(\omega)\}_{\ell=1}^{|J|}, \quad \zeta_\ell(\omega) = p_{j'}(\omega), \text{ where } j' = \text{argmax}\{\{p_j(\omega)|j \in J\}\backslash\{\zeta_{\ell'}(\omega)|1 \leq \ell' < \ell\}\}.$$

Tied terms are arranged arbitrarily in the above argmax function. Assume job j is positioned as k^{th} job in the sequence of jobs assigned to machine i. In the worst case, the jobs prior to job j are the ones with highest processing times, and job j has the highest processing time among remaining ones. Hence, summation of the first k elements of $\xi(\omega)$, $\sum_{\ell=1}^{k} \zeta_\ell(\omega)$, is an upper bound for $c_{ij}^k(\omega)$. Accordingly, the big-$\mathcal{M}$ in constraints (3.2b) can be replaced

by $\sum_{\ell=1}^{k}\zeta_\ell(\omega)$. This upper bound also works for the big-$\mathcal{M}$ in constraints (3.2c), since $\sum_{k'=1}^{k}\sum_{j'} p_{j'}(\omega)x_{ij'}^{k'} \leq \sum_{\ell=1}^{k}\zeta_\ell(\omega)$. In other words, $\sum_{\ell=1}^{k}\zeta_\ell(\omega)$ is large enough to ensure that constraint (3.2c) is relaxed in case $x_{ij}^{k}=0$. The obtained upper bound, $\sum_{\ell=1}^{k}\zeta_\ell(\omega)$, is also valid for the big-$\mathcal{M}$ in constraints (3.2d) and (3.2e) since the time a bed becomes available is the same as the time the patient's discharge is completed.

3.3 Solution Method

Because the model contains random variables and binary decisions, the deterministic equivalent is computationally challenging to solve in reasonable time for large-scale instances. Decomposition methods are often promising methods to overcome this difficulty. These methods decompose the deterministic equivalent problem (DEP) into a master problem and a subproblem. The master problem of the IDP model is presented in the following.

$$
\begin{aligned}
\text{(Master problem)}: \quad & \text{Min} \quad \sum_i\sum_j\sum_k \eta\alpha_j^k x_{ij}^k + \theta \\
& \text{s.t. } (3.1\text{b}) - (3.1\text{h}) \\
& {\boldsymbol{\beta}_t^h}^\top X + {\boldsymbol{\beta'}_t^h}^\top U + \theta \geq \gamma_t, \quad t=1,\cdots,h \qquad (3.3)\\
& x_{ij}^k, u_{im}^k \in \{0,1\}, \quad \forall i,j,k,m.
\end{aligned}
$$

Constraint (3.3) represents Benders-type optimality cuts generated in the first h iterations of the algorithm to approximate the expected recourse cost (θ), where ${\boldsymbol{\beta}_t^h}^\top X = \sum_i\sum_j\sum_k \beta_{ijt}^{kh}x_{ij}^k$ and ${\boldsymbol{\beta'}_t^h}^\top U = \sum_i\sum_m\sum_k \beta'^{kh}_{imt}u_{im}^k$. The $\boldsymbol{\beta}_t^h = (\beta_{ijt}^{kh} : \forall i,j,k)$ and $\boldsymbol{\beta'}_t^h = (\beta'^{kh}_{imt} : \forall i,m,k)$ are respectively $(ijk)\times 1$ and $(imk)\times 1$ vectors, and are known as cut coefficients for the t^{th} cutting plane in iteration h. The γ_t is cut constant. Details on calculating cut coefficients and the cut constant could be found under description of the stochastic decomposition algorithm in Section 3.3.1. Since the two-stage formulation of the IDP problem has relatively complete recourse (subproblem is feasible for every first-stage solution), there is no need to add feasibility cuts to the master problem. The subproblem for each realization

ω is the second-stage model presented by (3.2a)-(3.2f).

In this study, two decomposition algorithms including L-shaped and stochastic decomposition (SD) are implemented to solve the IDP model. These decomposition algorithms have three major steps repeated in every iteration until the algorithm converges or a stopping criteria is met. The steps include: (1) solving the master problem; (2) plugging in the obtained solution into the subproblem and solving that; (3) generating cuts and updating the master problem. L-shaped and SD are different in sub-steps of these major steps and also in convergence.

The widely used L-shaped algorithm solves the subproblem for all scenarios in every iteration. Using the optimal dual solution of the subproblem, a cut is generated and added to the master problem to improve approximation of the expected future cost. In terms of convergence, it is proved that L-shaped algorithm converges to optimal solution in finite number of iterations. However, the L-shaped method can be computationally burdensome for problems with a large number of scenarios. For more details on the L-shaped algorithm, see [75, 76, 77].

Unlike the L-shaped method, stochastic decomposition (SD) algorithm, developed by [78], solves the subproblem for only one randomly selected scenario in every iteration. This is the most important advantage of SD which significantly reduces the computational effort in generating a new cut. The algorithm then generates a new cut and adds that to the master problem. Another distinction between SD and L-shaped is that SD updates all previously generated cuts in every iteration. One more difference is that SD converges to an optimal solution asymptotically. By setting appropriate terminating criteria based on the instance characteristics, SD can find near optimal solutions. Details on terminating criteria for the IDP problem instances are discussed in Section 3.4.2. We refer interested readers to [79] for more details on SD.

We compare these two decomposition algorithms with a shortest expected processing time (SEPT) heuristic. It is the stochastic counterpart of the shortest processing time first

policy which is optimal for deterministic parallel machine scheduling. We modify the SEPT in order to account for bed request assignments in IDP problem. The idea is to sequence discharges with shorter expected processing times prior to longer ones, and to satisfy earlier bed requests with earlier discharges. As expected, computation time for SEPT is very short (around a few seconds), but the quality of solution should be investigated. The algorithm is discussed in details in Section 3.3.2.

3.3.1 Stochastic Decomposition Algorithm

Steps of the SD algorithm for solving the IDP model are presented below. In the algorithm presentation, $r^h - T^h X^h - T'^h U^h$ is the matrix format of right-hand sides of constraints (3.1b) − (3.1h) for the scenario generated in iteration h of the algorithm. Other notation is as follows.

h	Iteration number
V_h	Set of optimal dual solutions of the subproblem found up to iteration h
θ	Approximation of the expected recourse cost
L	Lower bound for the expected recourse cost
i_h	Iteration at which incumbent solution $\{\bar{X}^h, \bar{U}^h\}$ is found
Π	Set of feasible solutions of the dual problem of the subproblem
q	Fixed parameter used in updating the incumbent solution
γ_t^h	Cut constant of the t^{th} cutting plane in iteration h
$\boldsymbol{\beta}_t^h$	Vector of coefficient of X for the t^{th} cutting plane in iteration h, $\boldsymbol{\beta}_t^h = (\beta_{ijt}^{kh} : \forall i, j, k)$
$\boldsymbol{\beta'}_t^h$	Vector of coefficient of U for the t^{th} cutting plane in iteration h, $\boldsymbol{\beta'}_t^h = (\beta'_{imt}^{kh} : \forall i, m, k)$
$\{X^h, U^h\}$	Optimal solution of the master problem in iteration h
$\{\bar{X}^h, \bar{U}^h\}$	Incumbent solution in iteration h

Stochastic Decomposition Algorithm

Step 0. Initialization

Set $h \leftarrow 0$, $V_0 = \emptyset$, $\theta = -\infty$. Choose $q \in (0,1)$ and L.

Find an initial solution $\{X^1, U^1\}$. Set incumbent solution $\{\bar{X}^0, \bar{U}^0\} \leftarrow \{X^1, U^1\}$, $i_0 \leftarrow 0$.

Step 1. Generate ω^h

$h \leftarrow h + 1$. Randomly select a scenario ω^h from the set of scenarios for random variable $\tilde{\omega}$, independent of previously selected scenarios.

Step 2. Update V_h, generate the cut, and define $f_h(x)$

a. Update V_h

Plug $\{X^h, U^h\}$ into the subproblem and find an optimal dual solution to that; $\pi(X^h, U^h, \omega^h) \in \text{argmax}\{\pi(r^h - T^h X^h - T'^h U^h) | \pi \in \Pi\}$.

Do the same using the incumbent solution; $\pi(\bar{X}^{h-1}, \bar{U}^{h-1}, \omega^h) \in \text{argmax}\{\pi(r^h - T^h \bar{X}^{h-1} - T'^h \bar{U}^{h-1}) | \pi \in \Pi\}$.

Update V_h; $V_h \leftarrow V_{h-1} \cup \{\pi(X^h, U^h, \omega^h), \pi(\bar{X}^{h-1}, \bar{U}^{h-1}, \omega^h)\}$.

b. Calculate cut constant γ_h^h and cut coefficients $\boldsymbol{\beta}_h^h$ and $\boldsymbol{\beta'}_h^h$

$\gamma_h^h + \boldsymbol{\beta}_h^h X + \boldsymbol{\beta'}_h^h U = \frac{1}{h} \sum_{t=1}^{h} \pi_t^h (r^t - T^t X - T'^t X)$, where $\pi_t^h \in \text{argmax}\{\pi(r^t - T^t X^h - T'^t U^h) | \pi \in V_h\}$.

c. Update coefficients of the cut generated in iteration i_{h-1}

$\gamma_{i_{h-1}}^h + \boldsymbol{\beta}_{i_{h-1}}^h X + \boldsymbol{\beta'}_{i_{h-1}}^h U = \frac{1}{h} \sum_{t=1}^{h} \bar{\pi}_t^h (r^t - T^t X - T'^t U)$,

where $\bar{\pi}_t^h \in \text{argmax}\{\pi(r^t - T^t \bar{X}^{h-1} - T'^t \bar{U}^{h-1}) | \pi \in V_h\}$.

d. Update coefficients of all remaining cuts

$\gamma_t^h \leftarrow \frac{h-1}{h}\gamma_t^{h-1} + \frac{1}{h}L, \quad \boldsymbol{\beta}_t^h \leftarrow \frac{h-1}{h}\boldsymbol{\beta}_t^{h-1}, \quad \boldsymbol{\beta'}_t^h \leftarrow \frac{h-1}{h}\boldsymbol{\beta'}_t^{h-1}$

$\forall t \in \{1, ..., h\} \backslash \{i_{h-1}, h\}.$

e. Define $f_h(X, U)$

$f_h(X, U) = \sum_i \sum_j \sum_k \eta \alpha_j^k x_{ij}^k + \max\{\gamma_t^h - \boldsymbol{\beta}_t^h X - \boldsymbol{\beta'}_t^h U | t \in \{1, .., h\}\}.$

Step 3. Incumbent test

If $f_h(X^h, U^h) - f_h(\bar{X}^{h-1}, \bar{U}^{h-1}) < q\{f_{h-1}(X^h, U^h) - f_{h-1}(\bar{X}^{h-1}, \bar{U}^{h-1})\}$,

$\{\bar{X}^h, \bar{U}^h\} \leftarrow \{X^h, U^h\},\ i_h \leftarrow h.$

else

$\{\bar{X}^h, \bar{U}^h\} \leftarrow \{\bar{X}^{h-1}, \bar{U}^{h-1}\}, i_h \leftarrow i_{h-1}.$

Step 4. Solve the master problem

Solve the master problem to find $\{X^{h+1}, U^{h+1}\}$, then return to step 1.

3.3.2 Shortest Expected Processing Time First Heuristic (SEPT)

The Shortest Processing Time first policy, referred to as the SPT policy, is optimal for deterministic parallel machine scheduling with summation of completion times over all the jobs as the objective function [80]. Assuming m parallel machines, the SPT policy assigns the job with the shortest processing time to the first machine as the first job in the sequence. The second shortest job is assigned to the second machine, and the m^{th} shortest job is assigned to the m^{th} machine. The $(m+1)^{th}$ shortest job follows the shortest job on the first machine, the $(m+2)^{th}$ shortest job follows the job with the second shortest processing time on the second machine, and so on.

Because the objective for which the SPT policy is optimal, minimizing the sum of completion times, is meaningful in the IDP problem, it is promising to test an appropriately modified version of the stochastic counterpart, the SEPT heuristic, which sequences jobs by their expected processing times [80]. In the version presented below, patients are assigned

by the shortest expected processing time, then bed requests are assigned by the earliest bed request arrival time. The intuition here is that earlier bed requests are most likely to be satisfied by earlier discharges. Note that the SEPT heuristic does not consider patient preferences.

Extended SEPT Heuristic

Step 1: Calculate the expected discharge processing (EDP) time for each patient and expected bed request (EBR) time for each bed request. Sort patients and requests from lowest EDP and EBR to highest, respectively.

Step 2: Do the following for the set of patients sorted based on their EDP times:

For k' from 1 to $|K|$

For j' from $(k'-1)|I|+1$ to $\min\{|J|, k'|I|\}$

Starting from first nurse, assign j'^{th} patient in the set to first available position

Step 3: Repeat step 2 for the set of bed requests sorted based on the EBR times:

For k' from 1 to $|K|$

For m' from $(k'-1)|I|+1$ to $\min\{|M|, k'|I|\}$

Starting from first nurse, assign m'^{th} request in the set to the first available position

---

3.4 Computational Results

This section presents computational results for the IDP model using data obtained from a large cancer hospital in Houston, TX, USA.

3.4.1 Case Study Data

This study focuses on a general internal medicine unit in a large cancer center in Houston, Texas. In this unit, mid-level physicians round every morning and write patient discharge plans a day ahead based on their assessments and discussions with patient physicians. The unit receives bed requests from ED, direct admissions, and transfers from other units, with the majority of requests being from ED. In the course of our visits to the hospital, our partners in the office of performance improvement helped us get familiar with the inpatient discharge process, and confirmed the patient flow and IU dynamics depicted in Figure 3.1 as sufficient for their unit.

This inpatient unit has 48 beds divided evenly into four pods. Each pod is staffed by four nurses leading to a 1:3 nurse:bed ratio. Data provided covered 365 days in 2015. The data consisted of 2963 RFD patients and 3547 bed requests. Each RFD patient and bed request had an associated discharge or arrival date and time stamp. The data was used to fit distributions on the number of discharges per day, the discharge processing times, the number of bed requests per day, and the bed request arrival times (see Table 3.4). The fitted distributions are straightforward except for bed request times, which require some explanation. Bed request times are converted to number of minutes with 8:00 AM as the start point. Negative arrival times are thus the ones received before 8:00 AM. These are well represented by a normal distribution with mean of 393 minutes (2:33 PM). In scenario generation, we reset requests with negative arrival times to zero (which means those requests are present at the beginning of the day). The target discharge time for the inpatient unit is 3:45 pm daily. Although the unit is staffed with 16 nurses (4 nurses per pod), all might not be available or required for discharge-related tasks, as they may be fully occupied by non-discharge related work. Hence, the number of nurses required for discharging patients must be estimated (the model does not assume using the full capacity of nurses for discharging patients). Note that each nurse is in charge of up to 3 patients, any number of which could be RFD. Thus, the data is first used to estimate $p(\mathcal{Z})$ for $\mathcal{Z} \in \{0, 1, 2, 3\}$, that is, the

probability distribution of the number of RFD patients cared for by each nurse. To generate a realization of $\mathcal{Z}$ for each nurse i (denoted as n_i), we apply the Roulette Wheel selection method [81], the steps of which are summarized below:

Step 1. Define $F_{\mathcal{Z}} = \sum_{\mathcal{Z}'=0}^{\mathcal{Z}} p(\mathcal{Z}') \quad \forall \mathcal{Z} \in \{0,1,2,3\}$, $F_{-1} = 0$

Step 2. Generate a random number $\mathcal{X} \in [0,1]$

Step 3. If $F_{\mathcal{Z}-1} \leq \mathcal{X} < F_{\mathcal{Z}} \ \ \forall \mathcal{Z} \in \{0,1,2\}$, then $n_i = \mathcal{Z}$

Step 4. If $F_{\mathcal{Z}-1} \leq \mathcal{X} \leq 1, \ \mathcal{Z} = 3$, then $n_i = \mathcal{Z}$

Then, using the argmin formula given in the row $\#ofNurses : |I|$ in Table 3.4, we compute the number of required nurses for discharging RFD patients. In addition, the

Table 3.4: Inpatient Unit Case Study

Parameters	**Distribution**
# of Patients: $\|J\|$	Poisson($\lambda = 9.75$)
# of Bed Requests: $\|M\|$	Poisson($\lambda = 7.37$)
# of Nurses: $\|I\|$	$\text{argmin}\{n \mid \sum_{i=1}^{n} n_i \geq \|J\|, n_i > 0\}$
# of Positions: $\|K\|$	$\|K\| \in \{\lceil \frac{\|J\|}{\|I\|} \rceil, ..., \|J\| - (\|I\| - 1)\}$
Target Discharge Time	3:45 PM
Random Variable	**Distribution**
Discharge Process Time	Gamma($\alpha = 1.73, \beta = 164.27$) $\mu = 284.2$ min, $\sigma = 216$ min
Bed Request Time	N($\mu = 393$ min (2:33 PM), $\sigma = 6.9$ hours)

maximum number of patients that can be assigned to a nurse, $|K|$, needs to be determined. A lower bound on the value of $|K|$ is $\lceil \frac{|J|}{|I|} \rceil$, which ensures there is an available position for each patient. However, setting $|K|$ to its lower bound may cut an optimal solution from the feasible region of the first-stage model. Fortunately, an upper bound for $|K|$ can also be

developed. Imagine a case where one patient is assigned to every nurse, and all remaining $|J|-|I|$ patients are assigned to one nurse. In that case, $|K|$ takes the upper bound value of $|J|-(|I|-1)$. Setting $|K|$ to its upper bound is conservative and increases the computational time. Hence, there is a trade-off between solution quality and computational time.

3.4.2 A Designed Experiment

This section presents the results of three experiments. Fifteen IDP problem instances of different sizes were generated and solved. Where possible, the DEP of the two-stage model for each instance was solved first using IBM ILOG CPLEX Optimization Studio *V12.6.3*. In doing so, we built the DEP as an IloModel object using a C++ application with CPLEX in Concert Technology which is a C++ library. Then we used an IloCplex object to read the DEP model, extract its data, and solve it. Two decomposition algorithms, L-shaped and SD, were then applied to every instance. These algorithms were also implemented using CPLEX with Concert Technology in a C++ project. All experiments were performed on *Intel Core-i7, 2.9 GHz, 8 GB of RAM* computer. Table 3.5 provides the set sizes for each instance. Instances S1-S5 are relatively small, M1-M5 are medium-sized, and L1-L5 are large. The same number of scenarios were generated for each instance size.

For all experiments except the last, we assumed $\eta = 0.1$. That is, a higher weight is assigned to the second-stage objective since, for high levels of utilization, discharge lateness and patient boarding can have more negative consequences than failing to meet patient preferences. For example, ambulance diversion (due to ED boarding and subsequent backup of patients in the waiting area) can lead to patient death. Assuming $\eta = 0.1$ means that minimizing discharge lateness and patient boarding is 10 times more important than meeting patient discharge time preferences. It is also assumed that all bed requests are equally important.

The average processing time over all patients is used to specify the penalty for violating patient preferences. Therefore, both first-stage and second-stage objective functions are time-based. For example, assume that the average expected processing time is ρ and patient

j prefers to be discharged early in the morning, the penalty of assigning this patient to each position in the sequence is as follows:

$$\alpha_j^1 = 0, \alpha_j^2 = \rho, \alpha_j^3 = 2\rho \cdots$$

That is, no penalty is incurred for assigning the patient to the first position. However, if assigned to the second position, patient j has to wait for the patient in the first position to be discharged. Hence, patient j will be discharged ρ units of time, on average, later than the preferred discharge time. Thus, the penalty of assigning this patient to the second position is set to ρ. The penalty of assigning the patient to the third position is 2ρ by the same argument.

Table 3.6 provides patient preferences (preferred sequence position) used in the experimentation. These were uniformly randomly generated for each of the 15 instances since no preferences were given in the data. Table 3.7 presents SD parameters for each instance size. These values were chosen based on the computational experiments section of [82] where SD is tested for standard instances, with different sizes, in the literature of the stochastic programming. Twenty replications of the SD algorithm were performed for each instance. For small- and medium-size instances, a maximum number of 200 and 300 iterations, respectively, were allowed in every replication of the SD algorithm. For large instances, the maximum number of iterations was 500. In addition, a minimum of 100 iterations was imposed to prevent premature termination. For instances S1-S5 and M1-M5, the incumbent solution was required to remain unchanged for at least 50 iterations before terminating the algorithm. For instances L1-L5, the incumbent solution needed to remain constant for at least 70 iterations. Finally, all algorithms were terminated after 4 hours.

Table 3.5: Set Sizes for Each Instance

Size	Instance	# of Patients	# of Nurses	# of Bed Requests	# of Positions	# of Scenarios
Small	S1	4	2	4	2	250
	S2	4	3	4	2	250
	S3	5	2	5	3	250
	S4	6	5	4	2	250
	S5	7	4	5	3	250
Medium	M1	9	6	5	3	500
	M2	10	4	5	5	500
	M3	9	4	7	4	500
	M4	10	6	6	3	500
	M5	9	4	9	4	500
Large	L1	10	4	7	5	750
	L2	10	5	8	4	750
	L3	12	7	7	4	750
	L4	13	6	10	5	750
	L5	12	7	12	4	750

Table 3.6: Preferred Positions by Patients

	Patient												
Instance	1	2	3	4	5	6	7	8	9	10	11	12	13
S1	1	1	1	2	-	-	-	-	-	-	-	-	-
S2	2	2	1	1	-	-	-	-	-	-	-	-	-
S3	2	3	1	3	1	-	-	-	-	-	-	-	-
S4	1	1	2	1	1	2	-	-	-	-	-	-	-
S5	2	3	3	2	3	2	2	-	-	-	-	-	-
M1	1	1	3	3	3	2	3	2	2	-	-	-	-
M2	1	5	4	2	5	1	5	5	3	3	-	-	-
M3	1	3	2	3	2	4	3	2	2	-	-	-	-
M4	2	3	3	3	3	1	3	2	3	1	-	-	-
M5	2	4	1	2	3	1	3	3	2	-	-	-	-
L1	4	4	5	2	5	1	2	1	3	5	-	-	-
L2	1	3	3	2	1	4	2	2	2	3	-	-	-
L3	4	3	4	3	2	4	4	3	3	1	1	3	-
L4	5	2	5	3	5	1	1	5	3	3	2	4	4
L5	3	2	2	2	4	4	2	3	4	2	4	4	-

Table 3.7: SD Parameters for Each Instance

Instance Size	# of Replications	Min # of Iterations	Max # of Iterations	Min # of Iterations Soln. Stays Unchanged
Small (S1-S5)	20	100	200	50
Medium (M1-M5)	20	100	300	50
Large (L1-L5)	20	100	500	70

3.4.2.1 Experiment 1: Solution Accuracy and Speed

In this experiment, DEPs of 15 instances were attempted using CPLEX. Then, L-shaped and SD were also applied. Table 3.8 summarizes the results. Time units are in seconds.

Table 3.8: DEP vs. L-shaped vs. SD

	DEP Solved by CPLEX			L-shaped				Average Results for SD		
Instance	Best Soln.	Gap (%)	CPU Time	LB	UB	# of Iter.	CPU Time	Obj. Value	CPU Time	Deviation from CPLEX
S1	1136.4	0	68.7	1136.4	1136.4	575	114.2	1137.6	191.1	0.1%
S2	822.7	0	392.6	822.7	822.7	3455	22140.4	845.05	277.3	2.7%
S3	1936.5	0	1386.6	-5314.3	1939.5	2085	14400	1945.3	440.1	0.4%
S4	662.5	34.3	14400	-10433.5	676.9	811	14400	694.2	952.9	4.8%
S5	1343.9	60.5	14400	-23425.5	1396.2	649	14400	1427.5	1208.5	6.2%
M1	1718.4	76.1	14400	-36773.4	1833.1	366	14400	1751.5	2021.6	1.9%
M2	3914.3	87.9	14400	-89581.2	3760.5	317	14400	3805.2	3185.7	-2.8%
M3	2482.5	82.1	14400	-61509.8	2812.8	286	14400	2565.3	4329.2	3.3%
M4	1993.8	71.3	14400	-55956.6	2530.2	169	14400	2240.4	4862.9	12.4%
M5	3316.8	86.4	14400	341.1	5096.8	853	14400	3452.6	5283.4	4.1%
L1	-	-	-	-106375	4177.3	248	14400	3891.2	6902.6	-
L2	-	-	-	-86328	2771.5	197	14400	2560.4	7623.9	-
L3	-	-	-	-98290.5	4269.2	98	14400	4437.8	8752.6	-
L4	-	-	-	-187691	7121.2	89	14400	6941.5	10324.5	-
L5	-	-	-	-196230	6023.7	88	14400	5743.1	11694.8	-

As expected, for small instances like S1, decomposition was not needed, since solving DEP using CPLEX was the most efficient approach. For S2, which is larger than S1, CPLEX solved in 6.5 minutes and SD in 4.6. Average deviation of the SD solution from optimal was 2.7%, but SD improves computation time by 29.4%. In contrast, the L-shaped method did not converge (for S2, it required over 6 hours to converge, after which the time budget of 4 hours was applied). For S3 instance, SD was the most efficient. Compared to DEP, SD improves computation time by 68.3% with an average deviation from optimal of 0.4%. After 4 hours of computation, the L-shaped algorithm failed to converge.

For the DEP of S4, CPLEX failed to find the optimal solution in 4 hours, and the objective value of the best integer solution found was 662.5 minutes. The L-shaped algorithm also failed to converge in 4 hours. However, SD solved S4 in 15.9 minutes with an average error of 4.8%. This error is small compared to the improvement (94.4%) made by SD in computational time. Similar performance was observed for S5. CPLEX failed to identify the optimal solution and the L-shaped algorithm did not converge within 4 hours, but SD solved S5 in about 20.1 minutes with an average error of 6.2% from the best CPLEX solution.

For medium-size instances (M1-M5), CPLEX did not solve the DEP to optimality in 4 hours and the L-shaped algorithm exhibited a very large gap between its upper and lower bounds (note that the quality of the solutions found before termination/convergence is not known). In contrast, in every case, SD terminated successfully and achieved an objective value close to that obtained by CPLEX after 4 hours of computation. SD decreased computation time by over 60% in all cases. For M2, the objective value obtained by SD in about 54 minutes was 2.8% lower than the best objective value found by CPLEX in 4 hours. CPLEX was unable to solve the DEP for any of the large instances (L1-L5) due to memory constraints. Further, the L-shaped algorithm did not converge in 4 hours, and the gap between the lower bound and the upper bound was large. SD solved L1 in about 2 hours, and L5 (the largest) in about 3.2 hours. Further, the SD objective value was better than the upper bound obtained by L-shaped.

3.4.2.2 Experiment 2: Impact of Scenarios

Table 3.9 presents the impact of varying the number of scenarios for M3 and L3 from 500-1500 in increments of 250. Note that the DEP of M3 was not solvable (due to memory) starting from 1000 scenarios, and L3 was not solvable for 500 and up.

The L-shaped algorithm does not converge for any scenario level of M3 or L3 within 4 hours. In contrast, SD solved all instances of M3 and L3 with an average computation time about 1.4 hours and 2.65 hours, respectively. For M3 with 500 and 750 scenarios, the SD solutions deviated from the best DEP solution by 3.3% and -0.3%, respectively. Note that in one case, L3 with 750 scenarios, the L-shaped upper bound was slightly better than the objective value found by SD.

3.4.2.3 Experiment 3: SEPT Performance

This section analyzes the performance of the SEPT heuristic method. SEPT is intuitively appealing and quick, which could make it ideal for healthcare applications, but it is not designed to handle patient preferences. This experiment varies η, the weight on preferences, from 0 to 1 for instance S3. Table 3.10 provides a summary of results.

As expected, SEPT computational time was extremely small, 7.9 seconds. For the S3-0 instance ($\eta = 0$), the SEPT solution had a value 5.2% greater than the optimal. This gap increases as patient preferences become more important, growing to 53.5% when $\eta = 1$ ($\eta = 1$ implies that patient preferences are as important as minimizing discharge lateness and patient boarding). In experiment 1, we observed that SD was within 0.4% of optimal for S3 (with $\eta = 0.1$), thus outperforming SEPT for this case, which is 7.7% greater than optimal. Additional experimentation, not presented here, indicates that as problem size grows, this objective value gap between SEPT and SD is within 10% except when patient preferences are important. Thus, if patient preferences are not important, SEPT performs reasonably well. But, when preferences are important, SEPT is not appropriate.

Table 3.9: Impact of Scenario Size on Performance of DEP, L-shaped, and SD

	DEP Solved by CPLEX			L-shaped				Average Results for SD			
Instance	Best Soln.	Gap	Time (sec)	LB	UB	# of Iter.	Time (sec)	Obj. Value	Time (sec)	Deviation from CPLEX	Max # of Iter.
M3_500	2482.5	82.1%	14400	-61509.8	2812.8	286	14400	2565.3	4329.2	3.3%	300
M3_750	2703.5	83.2%	14400	-74346.2	2887.8	211	14400	2634.4	4176.8	-0.3%	300
M3_1000	-	-	-	-73234.4	2918.5	294	14400	2704.7	5369.1	-	400
M3_1250	-	-	-	-77996.2	2819.2	277	14400	2682.8	5471.2	-	400
M3_1500	-	-	-	-85672.5	2997.8	233	14400	2739.6	5329.6	-	400
L3_500	-	-	-	-125766	4748.5	108	14400	4253.4	8591.2	-	500
L3_750	-	-	-	-98290.5	4269.2	98	14400	4437.8	8752.6	-	500
L3_1000	-	-	-	-110901	5136.3	82	14400	4833.2	10176.5	-	600
L3_1250	-	-	-	-108388	5003.2	76	14400	4697.5	9934.6	-	600
L3_1500	-	-	-	-109687	5171.3	61	14400	4526.1	10203.9	-	600

Table 3.10: SEPT vs DEP

	DEP Solved by CPLEX					SEPT Heuristic				
	Opt. Obj. Value	Penalty	Discharge Lateness	Patient Boarding	CPU Time	Obj. Value	Penalty	Discharge Lateness	Patient Boarding	Gap from Opt.
S3-0	1848.6	0	711.9	1136.8	1535.9	1944.3	0	755.2	1189.1	5.2%
S3-0.1	1936.5	28.4	741.1	1167	1386.6	2086.4	142.1	755.2	1189.1	7.7%
S3-0.2	1964.9	56.8	741.1	1167	1269.4	2228.5	284.2	755.2	1189.1	13.4%
S3-0.3	1993.3	85.3	741.1	1167	1197.9	2370.6	426.3	755.2	1189.1	18.9%
S3-0.4	2021.7	113.7	741.1	1167	834.1	2512.7	568.4	755.2	1189.1	24.3%
S3-0.5	2050.2	142.1	741.1	1167	754.8	2654.8	710.5	755.2	1189.1	29.5%
S3-0.6	2078.6	170.5	741.1	1167	799.2	2796.9	852.6	755.2	1189.1	34.5%
S3-0.7	2107.0	198.9	741.1	1167	618.1	2939.0	994.7	755.2	1189.1	39.5%
S3-0.8	2135.4	227.3	741.1	1167	654.6	3081.1	1136.8	755.2	1189.1	44.3%
S3-0.9	2163.8	255.8	741.1	1167	914.4	3223.2	1278.9	755.2	1189.1	48.9%
S3-1	2192.3	284.2	741.1	1167	699.9	3365.3	1420.9	755.2	1189.1	53.5%

3.4.3 Benchmarking with Current Practice

This section presents experiment results used to benchmark the performance of two-stage IDP solutions with current practice. A discrete-event simulation model was developed (Arena, version 15) to compare the two-stage solutions with simple *ad hoc* methods used in practice. The simulation implemented a given schedule and followed the prescribed assignments and sequences of patients to nurses and bed assignments to available beds. Processing times and bed request arrival times were randomly selected during each simulation run. Consistency tests were performed to determine whether metrics computed by the simulator for a given schedule were similar to metrics computed by the two-stage model for the same schedule. This was deemed necessary to verify that the simulator dynamics accurately encompassed the more limited dynamics captured by the IDP model.

Due to space limitations, only the verification results for S3 are presented (other instances were tested with similar results). Table 3.11 provides the S3 configuration used. The first two patients were medical patients and the remaining three were surgical. Two bed requests came from the general internal medicine (GIM) unit, two from PACU, and one from ED. Table 3.11 provides the distributions for RFD patient processing times and bed request arrival times (fitted from the data).

Table 3.11: Simulation Parameters

Bed Request Arrival Times		RFD Patient Discharge Processing Times	
Source	Distribution	Patient Type	Distribution
GIM	N($\mu = 940, \sigma = 500$)	Medical	Gamma(1.73,201) $\mu = 347.7, \sigma = 264.4$
PACU	N($\mu = 750, \sigma = 120$)	Surgical	Gamma(1.73,140) $\mu = 242.4, \sigma = 184.1$
ED	N($\mu = 980, \sigma = 600$)		

Table 3.12 reports optimization and simulation metrics. The penalty for simulation is greater than the optimization model because the simulation computes penalties based on actual discharge time rather than using the approximate position-based method of Section 3.4.2. The discharge lateness and patient boarding values of the two-stage model are contained in the 95% CI of the simulation model. Thus, results indicate that simulator and two-stage model are reasonably consistent.

Table 3.12: Simulation Verification (Values in Minutes)

Performance Metrics	2-Stage Model	Simulation	
		Mean	95% CI
Preference Penalty	28.42	126.94	(122.59, 131.29)
Disch. Lateness	734.79	745.42	(686.07, 804.77)
Pat. Boarding	1020.28	1030.00	(958.61, 1101.39)
Objective	1783.49	1902.36	(1776.50, 2028.22)

To benchmark current practice, the simulator was used to compare schedules created with the two-stage approach and a simple rule easily applied in practice. Note that patient discharges are often assigned to nurses based on external reasons (e.g. location). The simulator implements this as a random assignment, then follows a simple sequencing rule. With the *TimePref* rule, a nurse will sequence patient discharges based on type (surgical before medical), since surgical patients are often less ill and have simpler discharge processes. Priority among patients of the same type is determined by patient preference. Any ties are resolved randomly. Further, bed requests are allocated on a first-come-first-serve basis.

Table 3.13 presents results. The two-stage model outperforms *TimePref* on all metrics (the total objective by approximately 30.7%). Further, none of the confidence intervals overlap, indicating that differences are statistically significant. Thus, it is reasonable to

conclude that the two-stage stochastic program can bring significant value to real hospital inpatient units.

Table 3.13: Comparison to Current Practice for S3 Instance (Values in Minutes)

Performance Metrics	Fixed Schedule		TimePref	
	Mean	95% CI	Mean	95% CI
Preference Penalty	126.94	(122.59, 131.29)	148.33	(142.63, 154.03)
Disch. Lateness	745.42	(686.07, 804.77)	989.81	(913.36, 1066.26)
Pat. Boarding	779.44	(707.18, 851.70)	1020.20	(928.39, 1112.01)
Objective	1651.80	(1525.12, 1778.48)	2158.34	(1992.38, 2324.30)

3.5 Conclusions and Future Research

We formulate the inpatient discharge planning (IDP) problem as a two-stage stochastic program with discharge processing time and bed request arrival time as two independent random variables. Our model optimizes the IDP problem from both the hospital and patient perspectives by minimizing upstream patient boarding and discharge lateness and by integrating patient preferences on discharge time of day. The IDP problem is solved for instances generated using real data from a large hospital in Texas. As the deterministic equivalent of the IDP two-stage model is not solvable in reasonable time, three solution approaches, L-shaped, stochastic decomposition (SD), and the shorted expected processing time first (SEPT) heuristic are developed. Among these, SD emerges as the only effective solution method. Overall, SD significantly improves computational time and achieves high solution quality in those cases that can be benchmarked.

In current practice, a variation of the SEPT heuristic was analyzed and results show the optimization approaches are considerably better. To implement the model, a scheduling tool

would be needed to interface with the model and underlying software. The nurse would use the model on a daily basis to make assignment and sequence decisions for IDP.

This work can be extended in several directions. The IDP problem can be formulated as a multi-stage stochastic program where every step of the discharge process, discussed in Section 3.1.1, is considered as a stage. Another extension is to formulate the IDP problem as a risk-averse two-stage stochastic program where the risk measure minimizes the positive deviation of the discharge time from the target time. Furthermore, bed requests can be prioritized base on their urgency. For example, one can assign the highest priority to bed requests from the emergency department. Non-discharge related tasks for nurses can be incorporated into the model explicitly. Nurse availability during the day can also be included as a random variable. In addition, nurse productivity can be integrated as a decreasing parameter with respect to fatigue level and time of the day. Most importantly, steps can be taken to apply the approach in practice so that real value may be realized.

Extreme events, such as a norovirus outbreak which results in longer hospital stays and more bed requests to IU, are not considered in our IDP problem. This work can be extended to manage such extreme events by taking capacity planning decisions such as surge IU beds, reserved nurses, and ambulance diversion into account. In this case, the objective function will be based on cost since a cost trade-off analysis is required. Thus, time-based delays should be translated into costs.

4. IMPACT OF POST-DISCHARGE PLACEMENT ON HOSPITAL CONGESTION AND COSTS

4.1 Problem Description

As explained in Chapter 1, while the majority of patients are discharged to home, some are transferred to post-acute care facilities (PACFs). Hospitals sometimes fail to find a suitable PACF for patients by the day they are medically ready to be discharged from the hospital. In such cases, patients remain in the IU, even though it is not medically necessary. In accordance with the research literature [7], we refer to these patients as alternate-level-of-care patients (ALC) and medically unnecessary days spent in the IU are referred to as ALC days. In addition to the ALC patients, some patients will stay in the IU longer than needed due to a lack of stable housing and/or decision-making capacity. Considering the high value of IU beds, this dissertation investigates the feasibility of creating a "post-discharge-unit" (PDU) for patients who are medically ready for discharge but are being delayed for some reason. More specifically, PDU capacity planning and cost-effectiveness issues are addressed, especially with respect to how the PDU influences upstream patient flow, e.g., ED congestion and hospital admission.

Figure 4.1 represents a schematic of the problem setting. The blue lines show the patient flows that we consider. The IU contains inpatients, some being ready for discharge (RFD). Most RFD patients will leave the hospital for home, so there is no barrier in discharging them, as long as the discharge process and transportation arrangements are complete. However, among patients who are medically ready for transition to the next level of care, some have a prearranged place in a PACF while others are still waiting for one to be located. Thus, a part of these patients can leave the hospital, as shown by a direct link from RFD to PACF in Figure 4.1. The remaining patients waiting for transition to a PACF (ALC patients) could be transferred to a PDU (if one existed and had available beds), stay there as long

as needed, and then be transferred. Figure 4.1 also shows the incoming patient flow from upstream units such as ED, ICU, and PACU. The IU receives bed requests from these units and admits patients to available beds. Although on any given day, RFD and ALC patients are known, the number and arrival time of bed requests throughout the day is not. This creates uncertainty in IU occupancy and thus uncertainty in required PDU capacity and cost-efficiency.

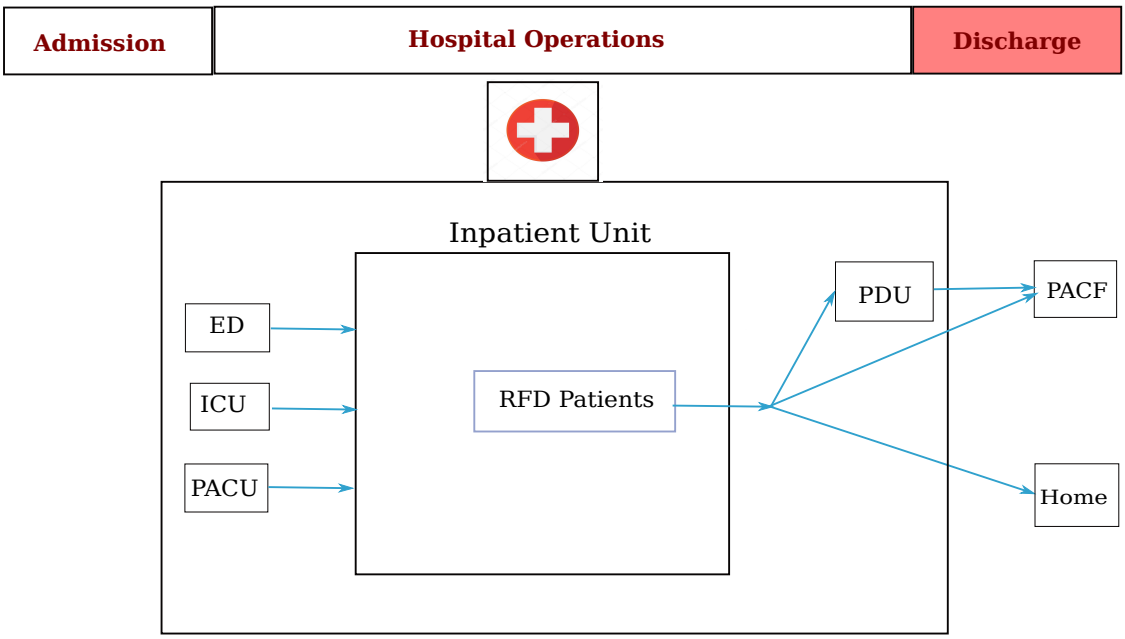


Figure 4.1: Patient Flow in Hospitals

This chapter models the patient flow throughout a hospital as a stochastic optimization program to find the optimal capacity of the PDU. A multi-dimensional random variable captures the uncertainty in demand for IU beds from upstream units. The objective is minimizing a hospital's total cost including fixed and operational costs of the PDU, the cost of ALC days, and IU bed request rejection costs. The latter is assumed to be the same as loss of revenue due to losing patients.

4.2 A Multistage Stochastic Programming Model

Most practical problems involve uncertainty which is typically revealed sequentially. In such problems, sequential decisions need to be implemented where the decision-making becomes more informed with evolving outcomes over time. An efficient approach to cope with sequentially evolving uncertainty within an optimization model is multistage stochastic programming. The general structure of a multistage stochastic program (MSP) is presented below assuming the stages are denoted by $t = 1, 2, \cdots, H$, $x^1 \in \mathcal{R}_+^{n_1}$, and $x^t(\omega^t) \in \mathcal{R}_+^{n_t}$ $\forall t$. The $\xi^t(\omega^t) = \{c^t(\omega^t), r^t(\omega^t), T^t(\omega^t)\}$ is a multi-dimensional random variable defined on the probability space (Ω, Ξ^t, P), where $\Xi^t \subset \Xi^{t+1}$ [76].

There is no uncertainty in the first stage meaning that the first-stage decision (here and now) is made before uncertainty is realized. Thus, all the data in the first stage including c^1, W^1, and r^1 are deterministic. The decisions in stages $t > 1$, $x^t(\omega^t)$, are called wait and see and depend on the realization of the random variable.

$$\text{Min } z = {c^1}^\top x^1 + \mathbb{E}_{\omega^2}[\text{Min } \ {c^2(\omega^2)}^\top x^2(\omega^2) + \cdots + \mathbb{E}_{\omega^H}[\text{Min } \ {c^H(\omega^H)}^\top x^H(\omega^H)] \cdots] \tag{4.1}$$

s.t.

$$W^1 x^1 \geq r^1 \tag{4.2}$$

$$T^2(\omega^2)x^1 + W^2 x^2(\omega^2) \geq r^2(\omega^2) \tag{4.3}$$

$$\vdots$$

$$T^H(\omega^H)x^{H-1} + W^H x^H(\omega^H) \geq r^H(\omega^H) \tag{4.4}$$

$$x^1 \geq 0, x^t(\omega^t) \geq 0, \quad t \in \{2, \cdots, H\}. \tag{4.5}$$

This dissertation develops an MSP to investigate the cost-effectiveness of a post-discharge unit and to determine its optimal capacity. The random variable is number of bed requests from upstream units, including the ED, ICU, PACU, transfers, etc. The demand for IU beds is stochastic for many reasons including the uncertain arrival rates of patients to a hospital

and service times in operating rooms and post-anesthesia care. Further, disastrous situations such as an outbreak of a pandemic disease could cause unexpected demand shocks.

The inpatient unit occupancy is directly impacted by the demand for its beds. If IU receives bed requests while operating on capacity, partially caused by ALC patients' non-medical stays, it has no choice other than rejecting the requests. This will incur the loss of revenue due to admission decline to IU, which could potentially prevent patient admission to upstream units, specifically the ED. If a post-discharge unit is available, the IU can transfer ALC patients to the PDU and admit new patients. This choice is not free of charge either as PDU is associated with a fixed and operational cost. However, it is expected that patient stay in PDU is less expensive than the IU. The objective function of our MSP model takes the trade-offs among these costs into account.

The decision on the PDU capacity is here and now. It means that a hospital needs to know the size when making the unit while the future demand for IU beds and its occupancy rate are unknown. Thus, PDU capacity is the first-stage decision. It is impacted by the IU occupancy rate, which itself depends on the uncertain demand for IU beds. The number of admissions to IU and transfers from IU to PDU depends on the IU occupancy level. All decisions which rely on the random variable (demand for IU beds/number of requests from upstream units for IU beds) are wait and see as they are made after realizing the uncertainty. In our multistage setting, the sequential wait and see decisions are the number of admissions to IU and transfers from the IU to PDU. These decisions should be made at different times during the day. To capture this variability, we break a day into multiple time intervals (stages). The following assumptions are used in developing the model:

- Patients are either discharged to their home or a PACF from IUs.
- Patients are admitted to a unit in the beginning of each stage.

Table 4.1 and Figure 4.2 show the notation for our MSP formulation.

Table 4.1: Notation

Sets	
J	Set of upstream units indexed by j
Parameters	
n^t	# of patients medically ready for transfer from IU to a PACF in stage t
h^t	# of patients that will be discharged from IU to home in stage t
r^t	Rate of patients for whom there is an available bed in a PACF in stage t
g	# of beds in IU
c	Fixed cost per unit capacity in PDU
c^w	Cost of non-medical stays in the IU per patient per night
c_j^r	Rejection cost of a bed request from source j
ω_t	Outcome of the random variable in stage t
$q_j^t(\xi_t)$	# of bed requests from source j under outcome ω_t
Decision Variables	
x	# of beds in the PDU
$p^t(\omega_t)$	# of patients transferred from IU to PDU at the end of stage t
$w^t(\omega_t)$	# of ALC patients in IU at the end of stage t
$v_j^t(\omega_t)$	# of bed requests from source j satisfied by IU in stage t
$u^t(\omega_t)$	IU Beds occupied by non-ALC patients at the end of stage t
$z^t(\omega_t)$	Beds occupied in PDU at the end of stage t

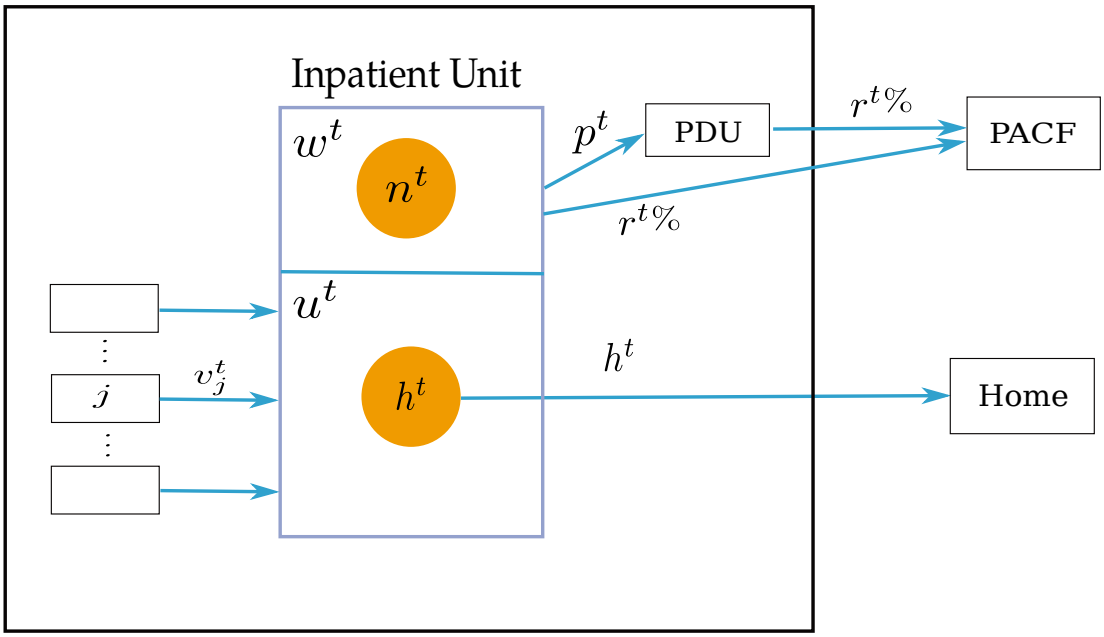


Figure 4.2: Notation

The MSP model for PDU capacity planning and cost-effectiveness evaluation is presented below. We name this model MSP-PDU, the first stage of which is defined in the following.

$$\begin{aligned} Q_1 = \quad & \min_{x} \quad cx + \mathbb{E}_{\tilde{\omega}_2}(Q_2(x, \omega_2)) \\ & \text{s.t.} \quad x \geq 0 \end{aligned} \tag{4.6}$$

The optimization model (4.7)-(4.14) presents the stochastic program for $t \in \{2, \cdots, T\}$.

$$Q_t(x, u^{t-1}, z^{t-1}, p^{t-1}, w^{t-1}, \omega_t) = \underset{w^t(\omega_t), v_j^t(\omega_t), p^t(\omega_t)}{\text{Min}} \quad c^w w^t(\omega_t) + \sum_j c_j^r (q_j^t(\omega_t) - v_j^t(\omega_t) + \mathbb{E}_{\tilde{\omega}_{t+1}}(Q_{t+1}(x, \omega_{t+1})) \tag{4.7}$$

s.t.

$$w^t(\omega_t) = (1 - r^t) w^{t-1} + n^t - p^t(\omega_t) \tag{4.8}$$

$$u^t(\omega_t) = u^{t-1} - n^t - h^t + \sum_j v_j^t(\omega_t) \tag{4.9}$$

$$u^t(\omega_t) + w^t(\omega_t) \leq g \tag{4.10}$$

$$z^t + p^t(\omega_t) \leq x \tag{4.11}$$

$$z^t = (1 - r^t) z^{t-1} + p^{t-1} \tag{4.12}$$

$$v_j^t(\omega_t) \leq q_j^t(\omega_t) \tag{4.13}$$

$$w^t(\omega_t), v_j^t(\omega_t), p^t(\omega_t) \geq 0 \quad \forall j, \tag{4.14}$$

where $Q_{T+1}(.) = 0$.

The first term in the objective function (4.6) is the fixed cost of making the PDU, and the second term is the expected future cost function. The present cost in the objective function (4.7) contains the cost of unnecessarily waiting in the IU and rejecting bed requests from upstream units. Constraint (4.8) captures the cumulative number of ALC patients waiting in the IU at the end of stage t. Constraint (4.9) is the balance constraint for the number of inpatient beds occupied by non-ALC patients at the end of stage t. Constraints (4.10) and (4.11) ensure that the number of occupied beds in the IU and PDU cannot exceed their capacities. Equation (4.12) obtains the state variable z^t. Constraint (4.13) ensures that IUs can only admit up to the number of bed requests. Constraint (4.14) imposes non-negativity restrictions on the decision variables in stage t. In equations (4.8) and (4.13), r^t is the ratio of ALC patients that can be transferred to a PACF immediately and it is equivalent to

$\frac{1}{\text{average ALC days}}$. For example, if the average ALC days is 5, 20% ALC patients are discharged every day.

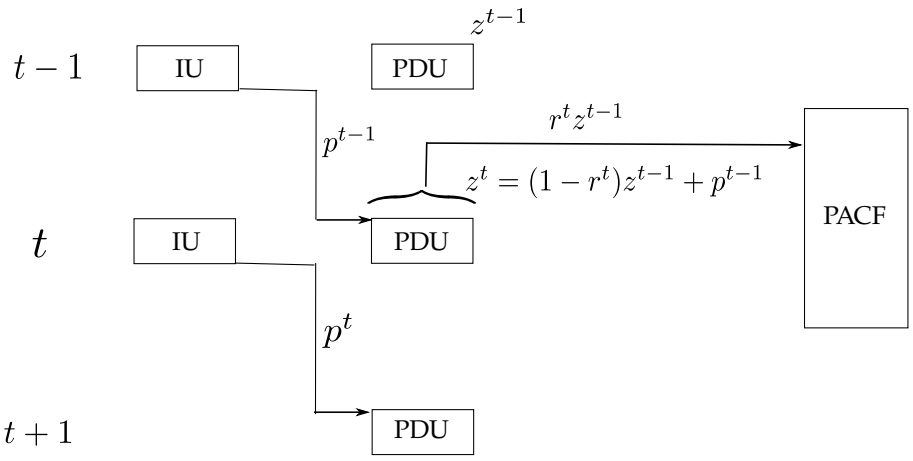


Figure 4.3: Patient Flow from IU to PDU

Figure 4.3 clarifies the patient flow from IU to PDU. Discharges from IU happen at the end of every stage, while patient admission to PDU is at the beginning of every stage. For example, assuming the current stage being interval 1 (6 AM-12 PM), p^t number of patients are discharged from the IU at the end of this stage (12 PM) and are admitted to the PDU at the beginning of the next stage (12 PM-6 PM). z^{t-1} is PDU occupancy at the end of stage $t-1$ and equivalently the beginning of stage t. r^t is the ratio of patients being discharged from PDU any time during stage t. This brings the total occupied beds in the end of stage t to $z^t = (1-r^t)z^{t-1} + p^{t-1}$, where p^{t-1} is the inflow to PDU in the beginning of stage t.

4.3 Stochastic Dual Dynamic Programming Algorithm

The size of a multistage stochastic program increases by number of stages and outcomes of the random variable per stage. This makes solving the deterministic equivalent problem computationally burdensome and almost impossible using optimization solvers such as CPLEX. Decomposition algorithms divide a multistage stochastic model into subproblems

(the stochastic optimization model in each stage) to overcome computational issues. The expected future cost function (the last term in Equations 4.6 and 4.7) in every subproblem is approximated using a piece-wise linear function through adding Benders-type cuts. Decomposition algorithms are iterative and the approximation of the future cost function is improved in every iteration by adding more informed cuts.

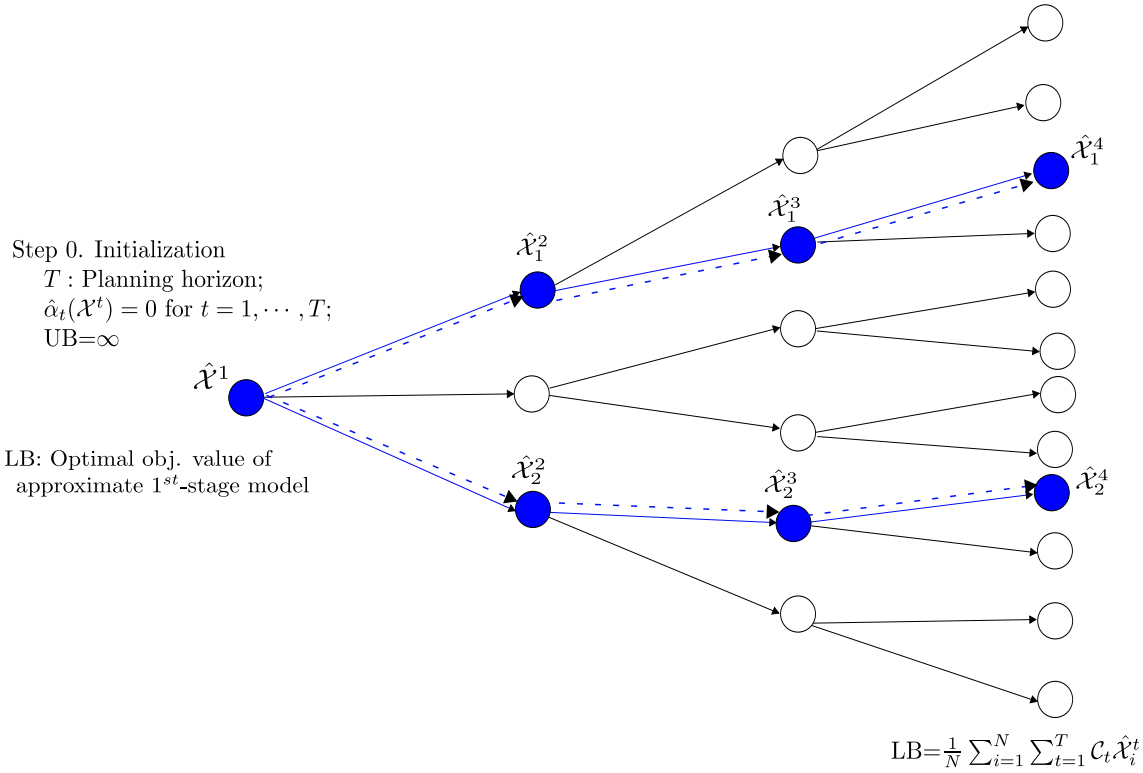


Figure 4.4: SDDP Forward Path

A standard decomposition method for solving multistage stochastic programs is stochastic dual dynamic programming (SDDP) algorithm which was initially developed to solve hydrothermal scheduling problems where the goal is to minimize expected costs. The algorithm has been widely adopted by academia and industry, and it is one of the most efficient techniques to deal with large-scale problems with hundreds of stages. There are several descriptions of SDDP in the literature, starting with the original work [83]. We base our

explanation in the description given in [84], which is fairly general. SDDP is proved to converge to an optimal solution in a finite number of iterations.

The SDDP has two major paths referred to as forward and backward. In the forward path, starting from the first stage, the subproblems are solved for N sample scenario paths. A scenario path is defined as a set containing an outcome of the random variable in every stage. At the end of each forward path, N trial solutions ($\hat{\mathcal{X}}_i^t, i \in \{1, \cdots, N\}$) are obtained for every stage. Figure 4.4 is a schematic of the SDDP forward path. This figure shows a scenario tree with 4 stages and 16 scenario paths. The two scenario paths highlighted in blue are sampled in this forward path ($N = 2$). Starting from the root node, the subproblem in every blue node is solved leading to two trial solution ($\hat{\mathcal{X}}_i^t, i = 1, 2$) in every stage. Note that the first stage is deterministic, so ($\hat{\mathcal{X}}_1^1 = \hat{\mathcal{X}}_2^1 = \hat{\mathcal{X}}^1$). A lower bound which is the optimal objective value of the first-stage model is updated in the beginning of each forward path. An upper bound, average of the total cost over sampled scenario paths, is obtained at the end of the forward path. The second major path of the SDDP is the backward path. As it is shown in Figure 4.5, starting from the leaf nodes, dual multipliers are obtained for all child nodes of a blue node. Then, a Benders type cut is generated using dual multipliers and added to the subproblem in the parent node. This continues until the algorithm reaches the root node. In words, every backward path improves the approximation of the stage-wise future cost functions by adding new cuts. The SDDP algorithm iterates between forward and backward paths until a stopping criteria (e.g., a maximum iteration limit) is met.

The subproblems for the MSP-PDU model and a detailed description of the steps of SDDP algorithm are presented in the following.

$$Q_1 = \quad \underset{x \geq 0}{\text{Min}} \quad cx + \hat{\alpha}_1, \tag{4.15}$$

$$\beta_1^h x + \hat{\alpha}_1 \geq \gamma_1^h, \tag{4.16}$$

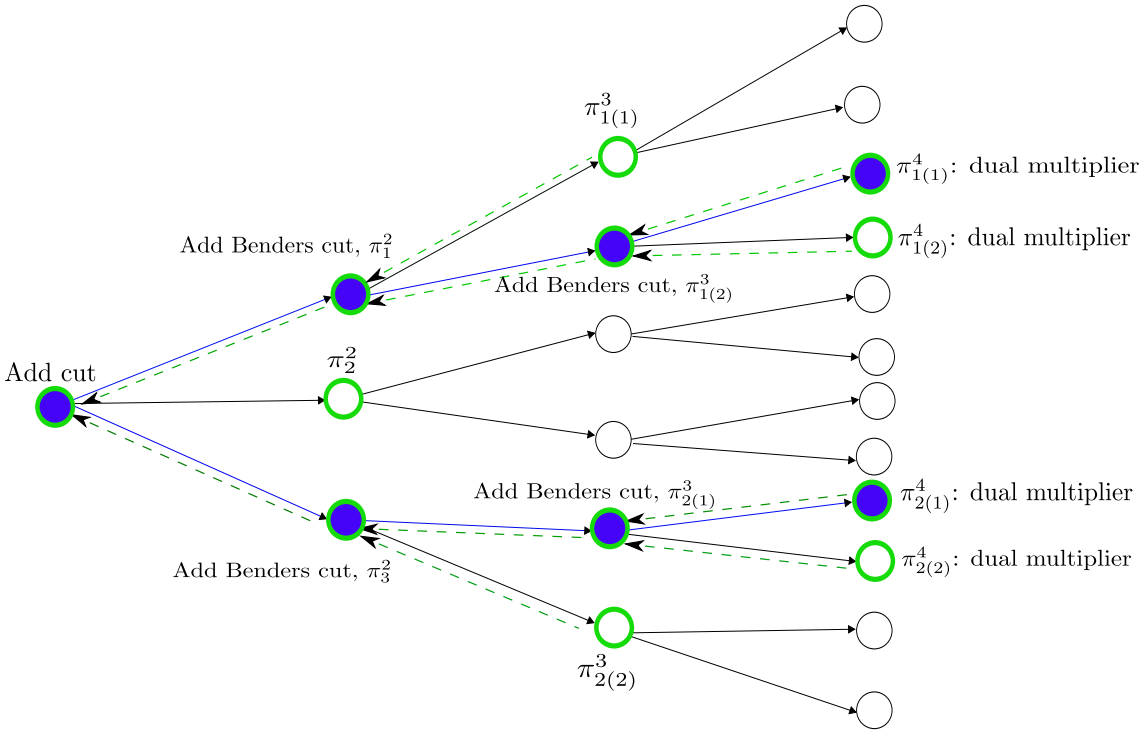


Figure 4.5: SDDP Backward Path

and the subproblems for $t \in \{2, \cdots, T\}$ are as follows. We name these subproblems PUD-SP.

$$Q_t(\mathcal{X}^{t-1}, \omega_t) = \quad \text{Min} \quad c^w w^t(\omega_t) + \sum_j c^r_j (q^t_j(\omega_t) - v^t_j(\omega_t)) + \hat{\alpha}_t \tag{4.17}$$

s.t.

$4.8 - 4.14$

$$\beta_t^{h\top} \mathcal{X}^t + \hat{\alpha}_t \geq \gamma^h_t, \tag{4.18}$$

assuming that $\mathcal{X}^{t-1} = (x, u^{t-1}, w^{t-1}, z^{t-1}, p^{t-1})$. Equation 4.16 and 4.18 are Benders-type cuts where $\beta_t^{h\top}$ and γ^h_t are the cut coefficient and constant in stage t, respectively.

SDDP Algorithm for MSPs

Step 0. Initialization.

Let T be the planning horizon; initialize $\hat{\alpha}_t(\mathcal{X}^t) = 0$ for $t = 2, ..., T$; upper bound $UB = \infty$.

Step (a): **Forward Path**

Solve problem 4.15.

Let $\hat{\mathcal{X}}^1$ be the optimal solution; initialize n trial decisions for the first-stage problem: $\hat{\mathcal{X}}_i^1 = \hat{\mathcal{X}}^1$ for $i = 1, ..., n$.

Update the lower bound (LB) which is the optimal objective value of problem 4.15. If $UB - LB \leq \epsilon$, stop; otherwise, go to the next step.

Step (b)

For $t \in \{2, ..., T\}$.

For each trial decision $\{\hat{\mathcal{X}}_i^{(t-1)}, i = 1, ..., n\}$.

Take a sample $\hat{\mathcal{V}}_t$ from the set of outcomes in stage t.

Solve the optimization problem PDU-SP in stage t using $\hat{\mathcal{V}}_t$.

Store the optimal solution as $\hat{\mathcal{X}}_i^t$.

Step (c): Update the upper bound

Calculate the upper bound as $UB = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \mathcal{C}_t^\top \hat{\mathcal{X}}_i^t$.

Step (d): **Backward Path**

For $t \in \{T, ..., 2\}$.

For each trial decision $\{\hat{\mathcal{X}}_i^{(t-1)}, i = 1, ..., n\}$.

For each outcome $\{\mathcal{V}_{tj}, j = 1, ..., m\}$.

Let π_{ij}^t be the dual multipliers associated with constraints of the PDU-SP in stage t at the optimal solution.

Calculate the expected vertex value: $\hat{\pi}_i^t = \sum_{j=1}^{m} p_j^t \pi_{ij}^t$

construct a supporting hyperplane of the approximate expected future cost function in stage $t-1, \hat{\alpha}_t$. Add the cut to problem PDU-SP in stage $t-1$.

Step (e).

Go to step (a).

4.4 Numerical Analysis

In this section, the MSP-PDU model is solved for several instances generated using the data from a large hospital in Texas. The case study is explained in Section 4.4.1. Some parameters in the model, such as the PDU operating and fixed costs, are estimated as the idea is new and is not yet implemented. Section 4.4.2 provides a detailed description of the parameter estimation. Numerical results are presented in Section 4.4.3.

4.4.1 Case Study

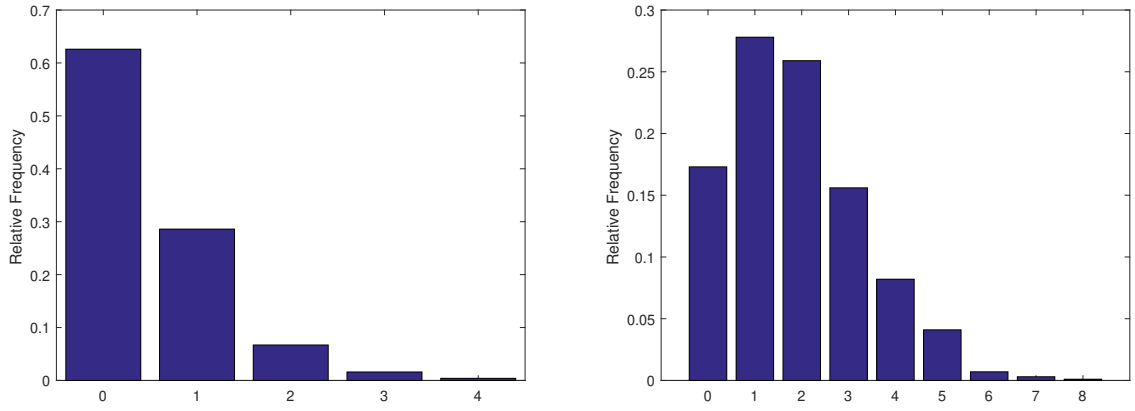
To analyze our model's performance, we use the data from a general internal medicine unit (G19) in a large cancer hospital in Houston, Texas. The unit accepts bed requests from different sources including ED, transfers from other inpatient units, direct admissions, and ICU. This unit includes four pods, each occupied by 12 beds and 4 nurses. Thus, the ratio of nurse to patient is 1:3. Patients in G19 are discharged to both home and PACFs. The ED in our focused hospital is frequently on diversion, and using beds in the hallways and chairs is sometimes inevitable.

Based on interviews with healthcare professionals in this hospital, transitions to post-acute care are challenging. Although a relatively small part (15%) of inpatients are transferred to PACFs in this hospital, some have extremely long wait times. In the best cases, patients are transferred to a PACF on the same day or the next day. However, in some cases, the patient is never transferred. The issue is usually finding a facility that fits the patient's needs rather than bed unavailability in PACFs. Most PACFs reject G19 patients for being

too medically complex (e.g., the need for high-cost medications, frequent radiations, and blood transfusion). Another issue in transferring patients to PACFs is the patient reluctance to leave the hospital. More than 50% of patients in G19 are not from Houston, and they come to our partner hospital to be treated by specific physicians, so they are not willing to leave the hospital.

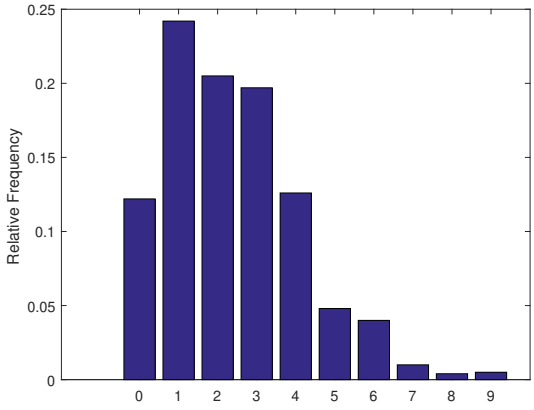
This hospital does not own any PACFs and do not contract with them. Texas medical center has plenty of PACFs in the greater Houston area. G19 patients are usually transferred to one of these facilities. Patient preferences play the primary role in the process of selecting a PACF. The hospital makes sure that patients are informed about Medicare rules and in-network and out of network facilities. The hospital also considers patients' geographic preferences. Some patients may need out of state referrals.

The data we received includes bed requests to G19 from different sources and their arrival times during 2014 and 2015. It also contains discharges and their associated times. Since admissions to and discharges from the G19 happen at different times during the day, we break a day into four intervals to better capture the dynamics involved. The intervals include 6 AM-12 PM, 12 PM-6 PM, 6 PM-12 AM, and 12 AM-6 AM, referred to as interval 1, 2, 3, and 4, respectively. The relative frequencies for the number of bed requests from different sources during each of these intervals are shown in Figures 4.6-4.9. Each figure includes four sub-figures (a)-(d), and they represent the number of bed requests during intervals 1-4, respectively. The figures confirm that ED is the primary source of G19 bed requests.

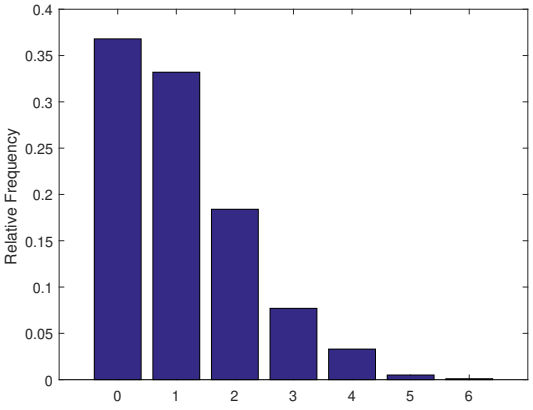


(a) Interval 1

(b) Interval 2



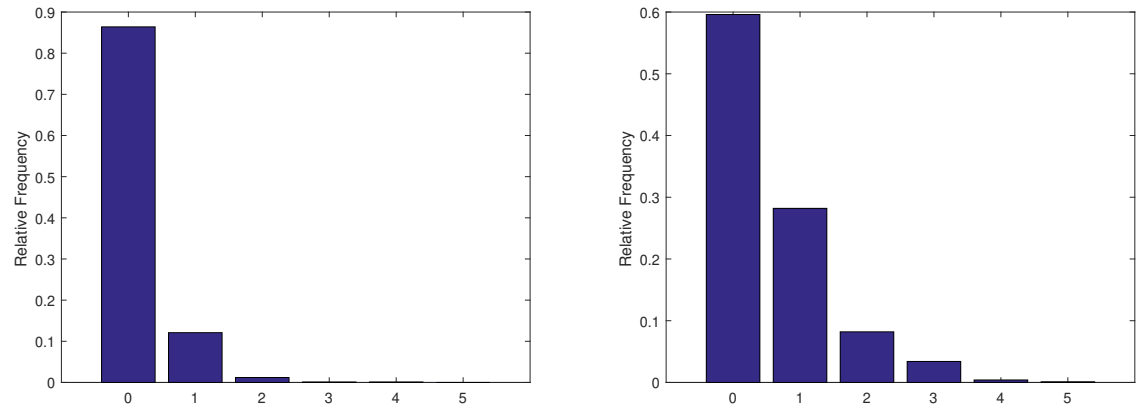
(c) Interval 3



(d) Interval 4

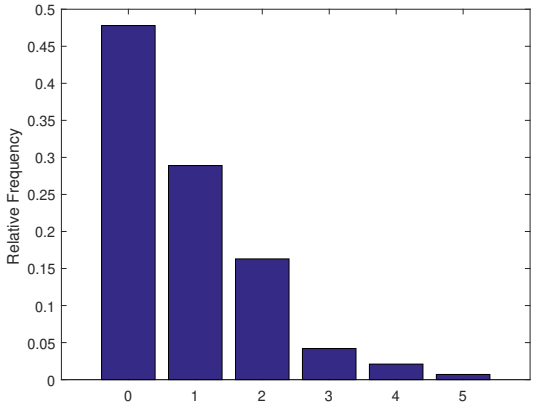
Figure 4.6: Number of Bed Requests from Emergency Department

As depicted in Figure 4.6, G19 receives a few bed requests from ED during interval 1 (no demand for IU beds has the highest frequency). Most requests arrive during intervals 2 and 3. The data shows that, in some days, the number of bed requests from ED was as highest as 8 and 9 during intervals 2 and 3, respectively. A few ED patients may need admission to the IU during interval 4.

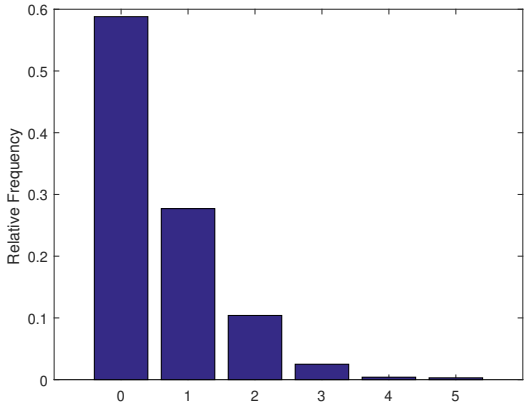


(a) Interval 1

(b) Interval 2



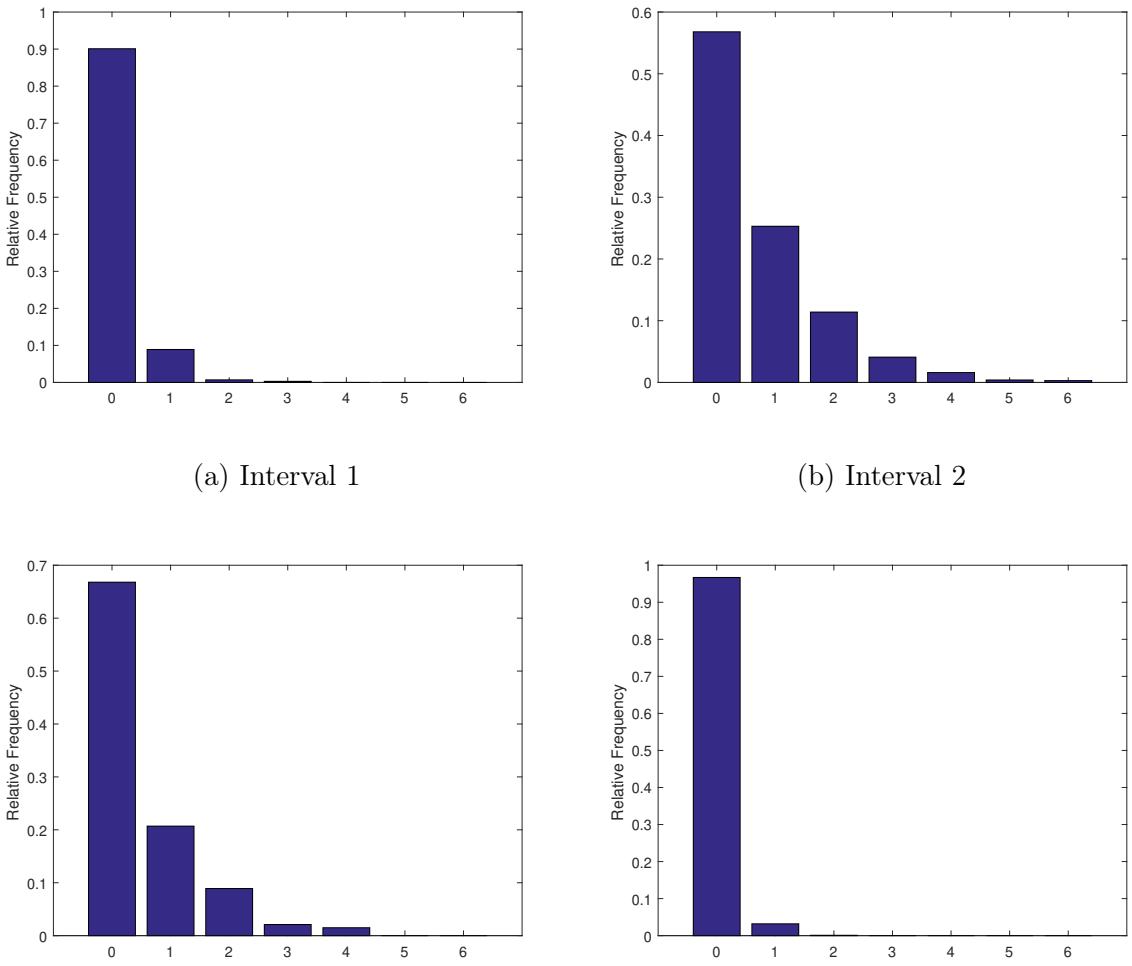
(c) Interval 3



(d) Interval 4

Figure 4.7: Number of Transfer Requests from Other Units to IU

Figures 4.7-4.9 show that transfers from other units are more frequent among the remaining sources of requests (direct admissions, transfers, and ICU). As explained earlier, direct admissions include patients being admitted from outside of the hospital directly to the IU. As Figure 4.8 shows, direct admissions, if any, are expected to arrive during intervals 2 and 3. Requests from ICU are negligible (Figure 4.9), so we will not consider them in our experiments.



(a) Interval 1

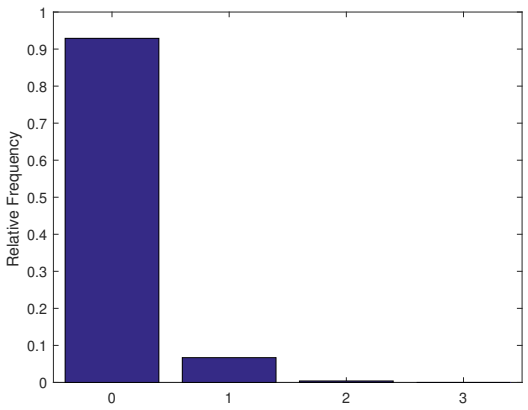
(b) Interval 2

(c) Interval 3

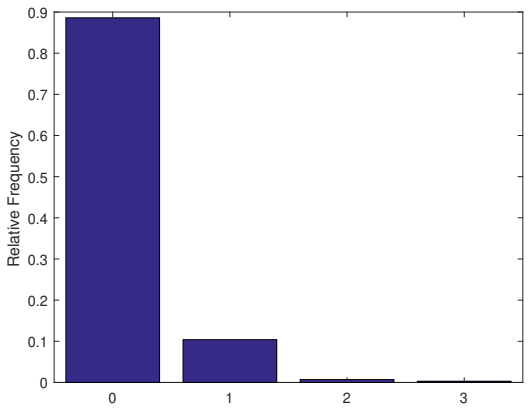
(d) Interval 4

Figure 4.8: Number of Direct Admission Requests to IU

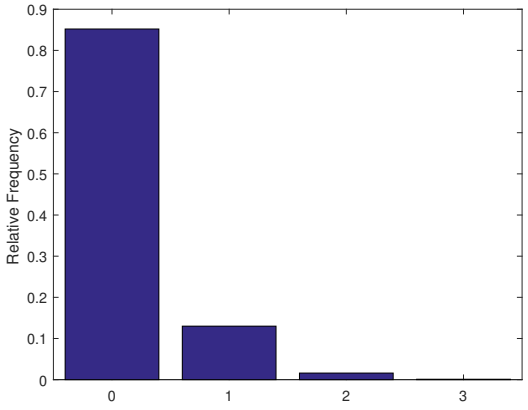
Table 4.2 shows the average number of bed requests from different sources per interval. This table confirms that majority of requests for G19 beds are made by the ED, especially during intervals 2 and 3.



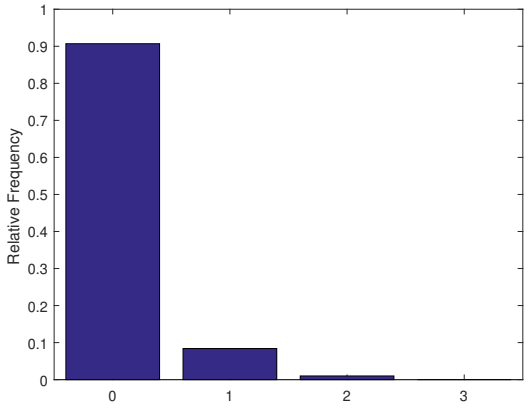
(a) Interval 1



(b) Interval 2



(c) Interval 3

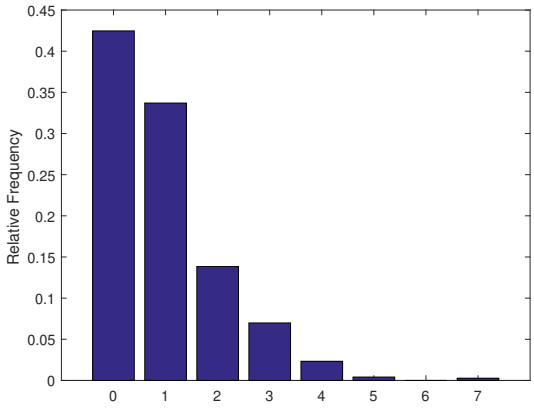


(d) Interval 4

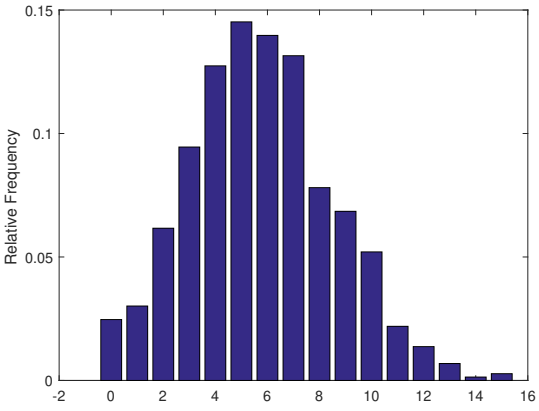
Figure 4.9: Number of Bed Requests from ICU

Table 4.2: Average Number of Bed Requests Per Interval

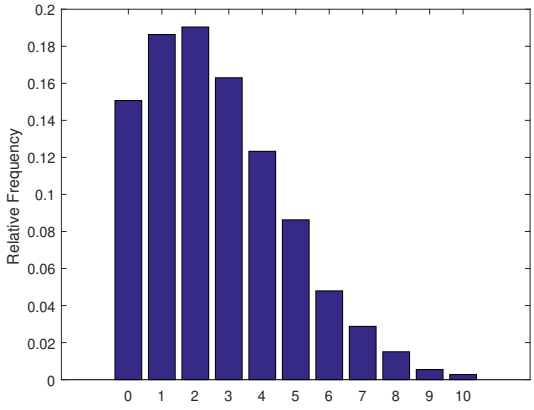
Source	Interval 1	Interval 2	Interval 3	Interval 4
ED	0.49	1.87	2.38	1.09
Transfer	0.15	0.57	0.86	0.59
Direct Admission	0.11	0.71	0.51	0.03
ICU	0.07	0.13	0.17	0.1



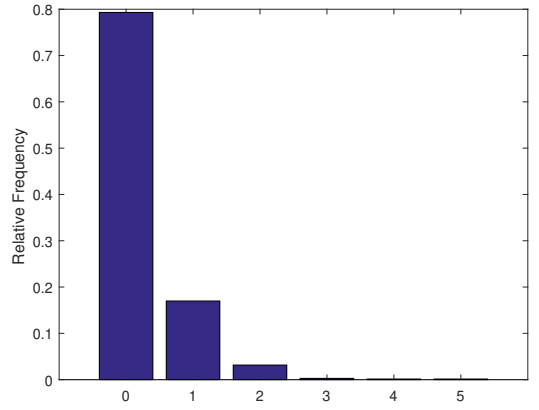
(a) Interval 1



(b) Interval 2



(c) Interval 3



(d) Interval 4

Figure 4.10: Total Number of Discharges (to both Home and PACFs)

Figure 4.10 presents the number of discharges in G19 during each interval. Most discharges happen during intervals 2 and 3 (12 PM-12 AM) with more frequency in interval 2. In G19, out of RFD patients, 15% will be discharged to the next level of care.

4.4.2 Parameter Description

This section explains parameter estimation, more specifically cost parameters, for the MSP-PDU model.

- *Cost of rejecting direct admission and transfer requests (per request)*: The inpatient unit loses revenue if refusing direct admissions and patient transfers. According to [85], the median revenue of a non-ED admission (transfers and direct admissions) is $4,118 per patient per day. The median length of stay for a non-ED admission is 3 days, so the total revenue is $12,354.

- *Rejection cost of requests for an IU bed made by ED (per request)*: Emergency department operates on or above the capacity in our partner hospital. Consequently, there is always a demand for ED beds. If the IU rejects a bed request, then ED should rely on its own resources to take care of arriving patients who are typically in urgent need of care. Declining admission of a new ED patient (ambulance diversion or patient walk-out) and using chairs and beds in hallways might be helpful. Such policies, especially declining new admissions, are costly for the hospital. Patients arriving by ambulance, specifically trauma patients, are among the main sources of profit. Ambulance diversion is a huge loss of revenue for a hospital. Admitting patients to hallways results in patient disappointment and ruins hospital's reputation, which is again a loss for the hospital. However, there is not a measure to quantify this loss.

  We estimate *the cost of rejecting a bed request from ED* as the expected revenue lost due to an ED patient left without being seen. ED patients are either outpatients or are admitted to the hospital. Expected revenue and percentage frequency for these two types of patients, based on the data from a large teaching hospital, are presented in Table 4.3 [85]. The weighted average of the median revenue, over patient types, is $2002, which we will use as the cost of rejecting a request for an IU bed made by the ED. Table 4.4 summarizes IU bed request rejection costs for different sources of request.

Table 4.3: ED Patients Data

Patient type	Outpatient	Admitted
Frequency (%)	78	22
Median revenue (per patient per day)	$647	$2,286
Median length of stay (days)	-	3

Table 4.4: Rejection Cost of IU Bed Requests from Different Sources (Per Request)

Emergency department	Direct admission	Transfer
$2002	$12,354	$12,354

- *Waiting cost of ALC patients in the IU (per patient per interval)*: The Kaiser Family Foundation (https://kff.org/health-costs/state-indicator/expenses-per-inpatient-day/) and [86] estimate that the average cost of a one-day hospital stay in the United States in 2017 was $2,424 ($101 per hour). We estimate an ALC patient's waiting cost in the IU during each interval as the multiplication of hourly IU-stay cost with the interval duration, which equals $606 for all intervals as they are of the same length.

- *Fixed cost of PDU (per unit capacity)*: We assume that hospitals have some space for being assigned to PDU, which means there are no land and major construction costs for building a PDU. However, the hospital may still need to pay for minor construction and renovation, basic medical equipment, and amenities. Making a small room larger, dividing a large room into two rooms, and adding a bathroom and hand-washing areas

are examples of minor constructions. Hospital furniture including beds, bedside tables, and chairs are also a part of the fixed cost. Some stand-by basic medical equipment such as blood pressure monitor and IV pumps are also needed for PDU. Amenities (e.g., TV) are another part of the fixed cost. We do a sensitivity analysis on the total fixed cost per unit capacity of the PDU. We assume a range of \$8,000-\$44,000 with increments of \$4,000 (equivalent to \$960K-\$5,280K (per unit capacity) in ten years). The fixed allocation for one month is similar to a mortgage payment. We consider the monthly fixed cost in our optimization model since we solve the model for 120 stages/30 days.

- *Operational cost of PDU (per patient per interval)*: In addition to the fixed cost of building a PDU, patient stay in this unit is associated with an operational cost. Nursing staff and social workers are the main sources of operational cost in the PDU. Patient supplies also contribute to this cost. As explained in Chapter 1, patients in the PDU do not require diagnostic medical equipment, and they need less nursing care than inpatients. Thus, we believe that operational cost in a PDU is less compared to an IU. Since a PDU has never been studied, and there is no related data in the literature, we quantify the PDU operational cost as a ratio of the waiting cost in IU. We consider a range of 50%, 60%, and 70% of IU waiting cost (\$606 per patient per interval) as the operational cost in PDU.

- *Average ALC days*: ALC days are the length of ALC patients' unnecessary stay in a hospital after being medically ready for discharge. The average ALC days are estimated at 4.8 [87]. However, to be conservative, we assume an average of 3 days in our initial experiments. We then compare the results with those obtained after increasing this parameter to 5.

4.4.3 Design of Experiments

This section studies the impact of variation on PDU fixed cost, PDU operational cost, the ratio of inpatients being transferred to PACFs, and average ALC days on PDU capacity. We use the SDDP package [88] in Julia *V1.0.5* and Gurobi as the solver in order for solving the MSP-PDU model.

4.4.3.1 Impact of PDU Costs on Its Capacity

Based on interviews with healthcare professionals in our partner hospital, 15% of inpatients in G19 are transferred to PACFs. This ratio is different depending on the type of inpatient unit. For instance, the ratio is significantly larger in inpatient units with stroke and mental health patients. Based on a study in a large acute hospital [87], 41.7% of stroke patients, 65.2% of patients with hip fractures, and 63.6% of patients with amputation are discharged to PACFs. Thus, we assume three hospital settings with 15%, 30%, and 60% inpatient transfer to post-acute care facilities named H1, H2, and H3, respectively. For each of these settings, we solve the MSP-PDU model to find the optimal PDU capacity for different variations on PDU fixed and operational costs.

Results are presented in Figures 4.11-4.13. Each figure shows the results for PDU capacity considering a range of monthly fixed costs ($8000-$44000 with increments of $4000) and operational costs (50%, 60%, and 70% of the waiting cost in IU). In this experiment, we assume that ALC patients wait an average of three days in IU before leaving the hospital. This assumption is to be conservative as the average ALC days is estimated at 4.8 days in the literature.

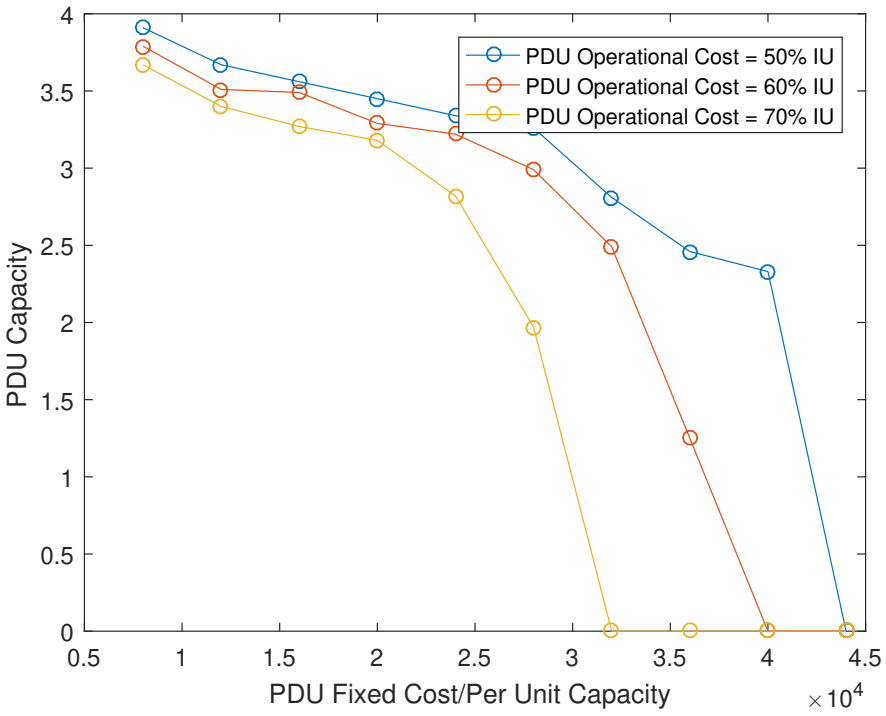


Figure 4.11: Hospital Setting with 15% Transfer to PACFs

Figure 4.11 shows the results for our focused hospital, where only 15% of inpatients are discharged to post-acute care facilities. The MSP-PDU model finds building a PDU cost-efficient even when both the ratio of transition to PACFs and average ALC days are reasonably small. As expected, the PDU capacity decreases by increasing its fixed and operational costs. However, the deviation among PDU sizes for different levels of the operational cost (specifically when the fixed cost is lower than \$28K) is minor. It is not cost-efficient to build a PDU if the fixed cost goes beyond \$44K.

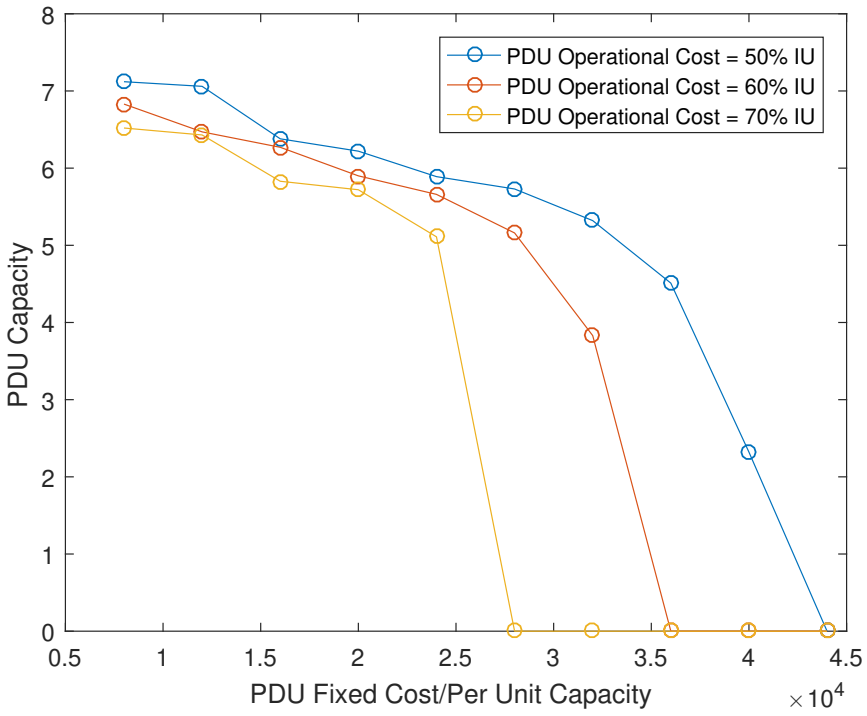


Figure 4.12: Hospital Setting with 30% Transfer to PACFs

Results for H2, where 30% of inpatients are transferred to PACFs, are presented in Figure 4.12. As expected, the larger the ALC patients population is, the bigger the PDU should be. In this setting, the largest PDU will have 7.35 beds. Note that the capacity of our partner hospital unit is 48, so a PDU with 7.35 beds is 15.3% of the IU size. This happens when the monthly fixed cost is \$8000, and the operational cost is 50% of the IU stay cost. Again, the deviations among the three plots are not significant. As the figure shows, a PDU is not beneficial for the monthly fixed cost greater than \$44K (per unit capacity).

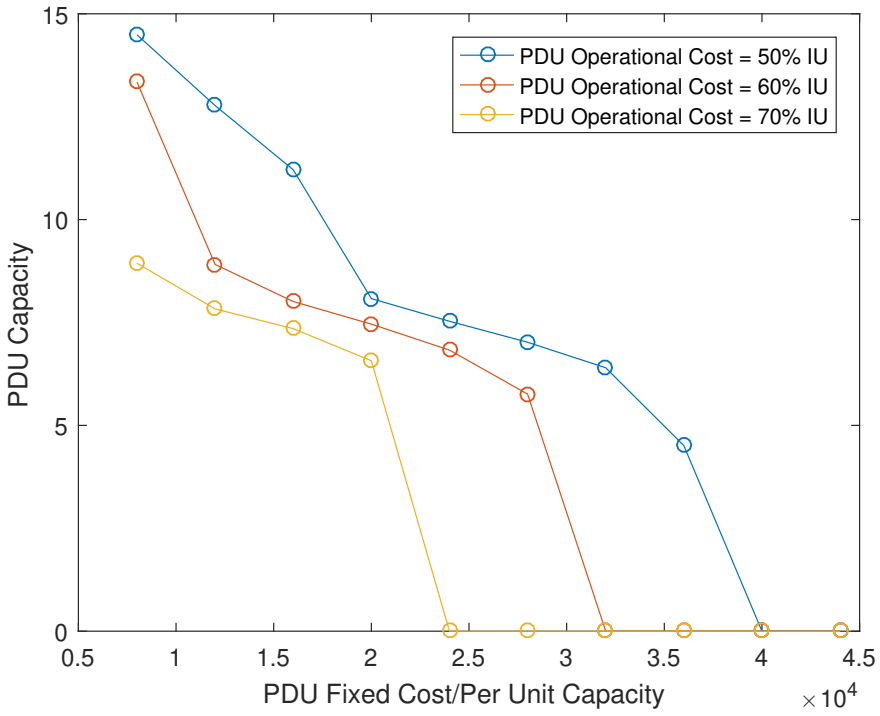


Figure 4.13: Hospital Setting with 60% Transfer to PACFs

Figure 4.13 shows the results for hospital setting H3. Compared to the two previous settings, the PDU capacity is significantly larger. The largest PDU among different combinations of the fixed and operational costs is 30% of the IU size, and the smallest one is 20% of it. Table 4.5 summarizes the results for the PDU capacity as a fraction of the IU size ($\frac{\text{PDU Capacity}}{\text{IU Capacity}} \times 100$). Given any level of the PDU operational cost, comparing the subtraction of subsequent rows in Table 4.5 shows that *H3 is more sensitive to increasing the fixed cost.* For example, assuming that the PDU operational cost is 50% of the IU's and increasing the fixed cost from \$16K to \$20K, the reduction in PDU capacity for H1, H2, and H3 is 0.2%, 0.3%, and 6.5% of the IU capacity, respectively.

Since H3 is more sensitive to the fixed cost changes, considering a fixed operational cost, the ratio $\frac{\text{PDU Capacity in H3}}{\text{PDU Capacity in H2}}$ decreases faster compared to $\frac{\text{PDU Capacity in H2}}{\text{PDU Capacity in H1}}$ by increasing the fixed cost. In fact, $\frac{\text{PDU Capacity in H2}}{\text{PDU Capacity in H1}} \approx 2$ stays almost constant. Doubling the transfer rate

to PACFs from 15% to 30% almost doubles the PDU size no matter what the fixed cost is. However, this is not true when the transfer rate increases from 30% to 60%. Further, Figure 4.13 is different than Figures 4.11 and 4.12 in the sense that the gap between the plots is larger. We use Figure 4.14 to elaborate this better.

Table 4.5: (PDU Capacity/IU Capacity)×100

Operational Cost	50%			60%			70%		
Fixed Cost ($)	H1	H2	H3	H1	H2	H3	H1	H2	H3
8K	8.1	14.8	30.2	7.9	14.2	27.8	7.6	13.6	18.6
12K	7.6	14.7	26.6	7.3	13.5	18.6	7.1	13.4	16.3
16K	7.4	13.3	23.4	7.3	13.1	16.7	6.8	12.1	15.3
20K	7.2	13.0	16.8	6.9	12.3	15.5	6.6	11.9	13.7
24K	7.0	12.3	15.7	6.7	11.8	14.2	5.9	10.6	0
28K	6.8	11.9	14.6	6.2	10.8	12.0	4.1	0	0
32K	5.9	11.1	13.3	5.2	8.0	0	0	0	0
36K	5.1	9.4	9.4	2.6	0	0	0	0	0
40K	4.9	5.4	0	0	0	0	0	0	0
44K	0	0	0	0	0	0	0	0	0

Figure 4.14 shows the percentage reduction in the PDU capacity for the highest operational cost (70% of IU stay) compared to the lowest (50% of IU stay) in each of the three hospital settings. More specifically, Figure 4.14 represents the absolute value of $\frac{\text{PDU size with 70\% IU cost - PDU size with 50\% IU cost}}{\text{PDU size with 50\% IU cost}}$ for each hospital setting. The brown line shows H3, where the majority of patients are transferred to PACFs, and they wait an average of 3 days before leaving the hospital. In this hospital setting, ALC patients are the primary

source of IU bed occupancy. Thus, the presence of a unit for ALC patients, which is less expensive compared to IU, is exciting. The PDU capacity in H3 with the operational cost being 50% of IU's and a monthly fixed cost of 8000 is 14.49. This is 30.2% of the IU size and is relatively large. However, as the PDU operational cost increases (gets closer to the inpatient cost), its advantage over the IU decreases. Increasing PDU operational cost to 70% causes a 38.3% reduction in capacity (from 14.49 to 8.94).

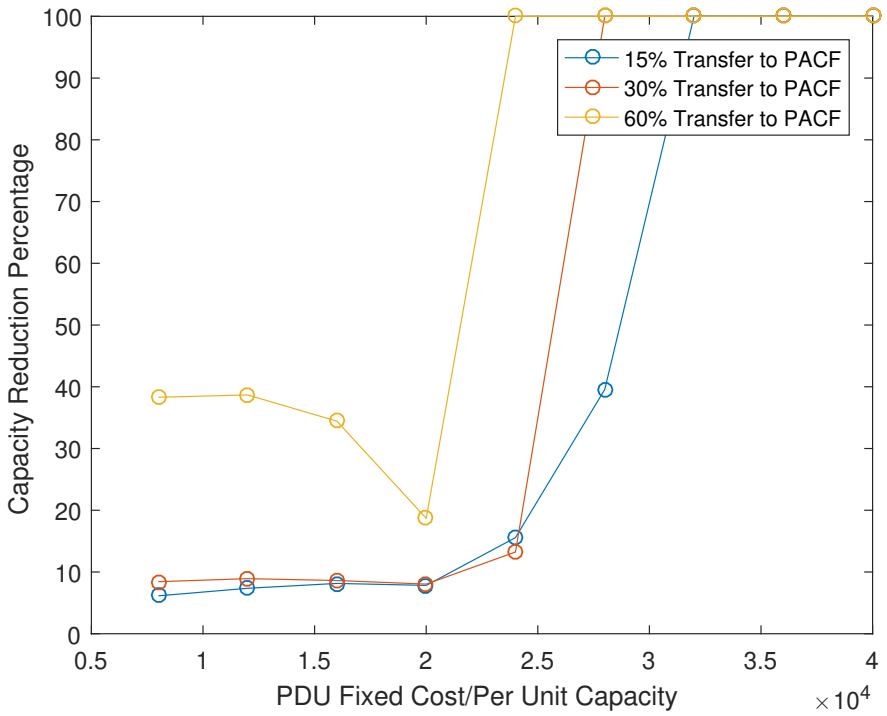


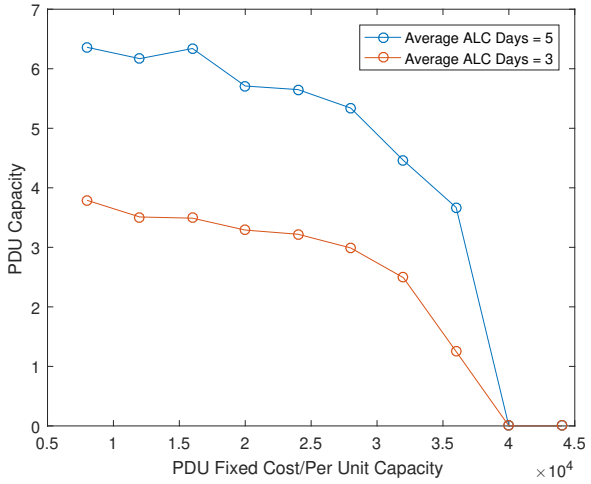
Figure 4.14: Reduction Percentage of PDU Capacity for Operational Cost 50% IU Cost vs 70%

For H1 and H2, the PDU, although still beneficial, is not as critical compared to H3. The largest PDUs have 7.12 and 3.91 beds (equivalent to 14.8% and 8.1% of IU capacity) for H2 and H1, respectively. If PDU operates on 70% of the IU cost, these values decrease to 6.52 and 3.67 (8.43% and 6.14% reduction). The reason is that the PDU for these two

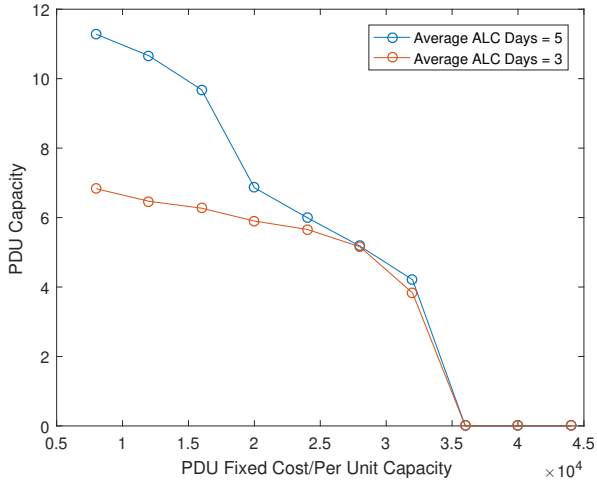
hospital settings is already relatively small. The smaller the PDU, the lower the impact of operational cost on the capacity is. Note that the reduction rate is lower for H1 compared to H2 (the blue line is almost always under the red line). *Overall, an important finding is that PDU capacity in hospitals with a more extensive transfer to PACFs is more sensitive to PDU costs including the operational and fixed costs.* Thus, such hospitals need to have a better estimate of these costs to design the most cost-efficient PDU.

4.4.3.2 Impact of Average ALC Days on PDU Capacity

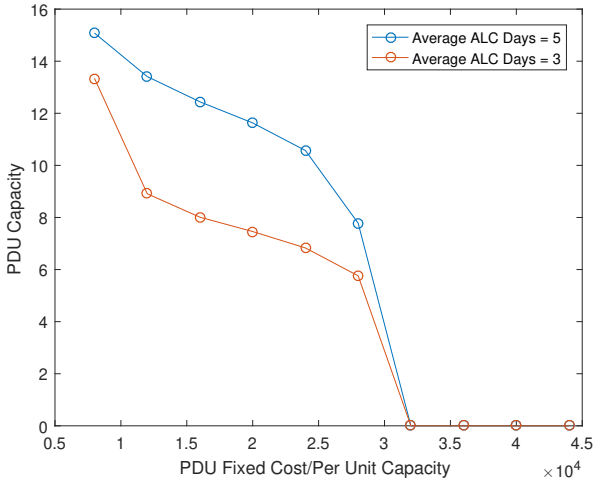
In the previous experiments, to be conservative, we assumed that the average ALC days is 3 although it is estimated at 4.8 based on the literature [87]. In this experiment, we increase the average ALC days to 5 and fix the PDU operational cost at 60% of IU stay cost. This experiment studies the impact of longer average ALC stays on the PDU capacity for the three hospital settings H1, H2, and H3. The results are presented in Figure 4.15. The figure shows that the PDU capacity increases with the average number of ALC days, which is reasonable. There is a large gap between the optimal PDU size for average ALC days being 3 vs. 5. This gap is above 50%, for most of the assumed fixed costs, in all three hospital settings. The gap closes by increasing the fixed cost. In other words, the impact of ALC days on the PDU capacity becomes less critical for higher fixed costs. In all three hospital settings, the blue and the red lines intersect on the x-axis (PDU size 0). It means that creating a PDU becomes inefficient beyond the same fixed cost, for both 3 and 5 average ALC days. Further, the fixed cost, beyond which creating a PDU becomes inefficient, decreases by moving from H1 (transfer rate 15%) towards H3 (transfer rate 60%).



(a) Results for H1



(b) Results for H2



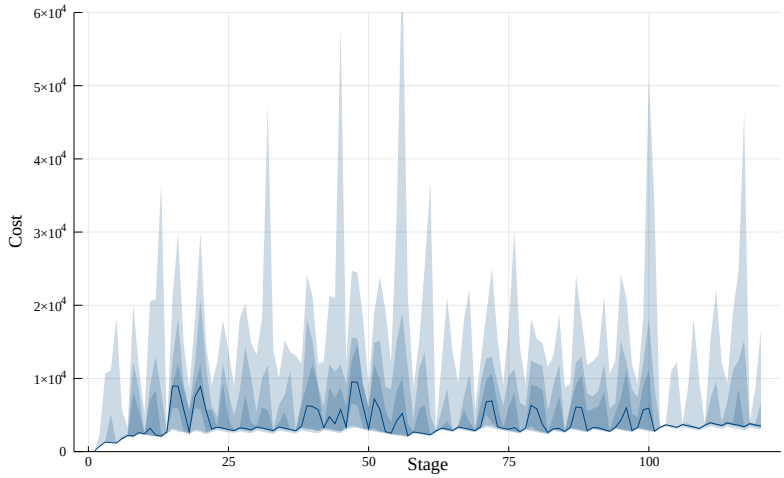
(c) Results for H3

Figure 4.15: Comparing the Results for Different Average ALC Days

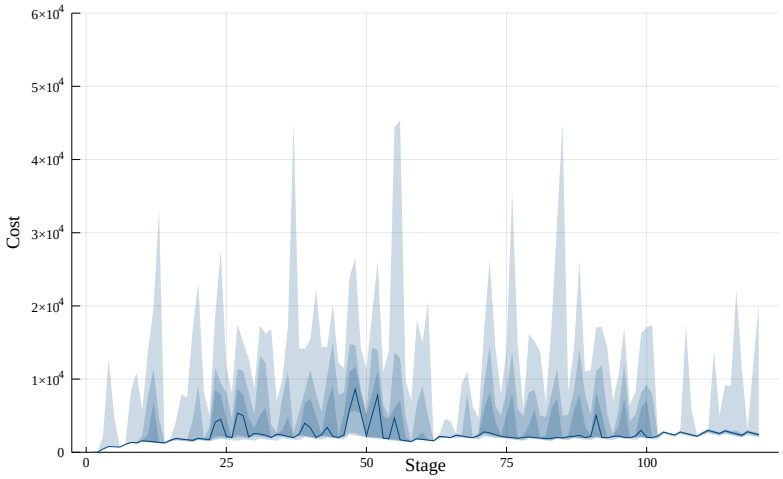
4.4.3.3 Benchmarking with Current Practice

In this section, we compare the current practice in our partner hospital with the optimal policy obtained from the MSP-PDU model. The average number of ALC days is assumed to be 5; the PDU fixed cost is \$16K per month, and its operational cost is fixed at 60% of the waiting cost at IU. The results are presented in Figures 4.16-4.23. The MSP-PDU model

finds the optimal PDU size at 5.87 for this hospital setting. To evaluate the current practice, we use our MSP-PDU model, enforcing the PDU capacity to 0. It means that we assume the hospital's performance is optimal, which is conservative.



(a) Current Practice

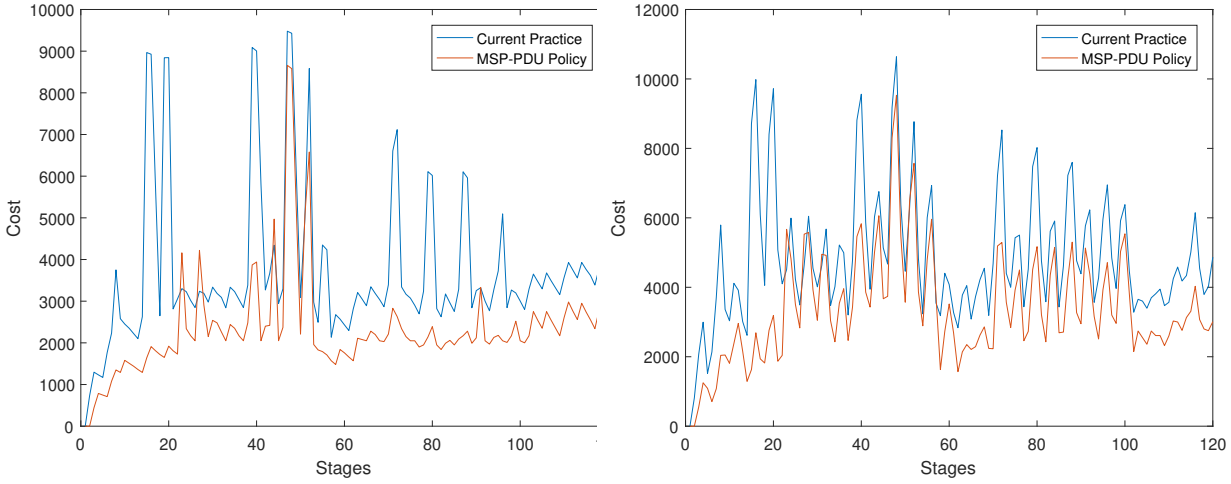


(b) MSP-PDU Policy

Figure 4.16: Current Practice Costs vs. the MSP-PDU Policy

Figure 4.16 compares the current practice vs. creating a PDU in terms of costs. Sub-figures (a) and (b) show the stage-wise immediate costs (excluding the initial stage) for the

current policy (no PDU) and MSP-PDU policy, respectively. Each figure represents the results of evaluating the policy for 5000 randomly selected scenarios. Three shades of color could be distinguished in these plots: from lighter to darker, colors indicate 0-100, 10-90, and 25-75 percentiles, respectively. The solid line in the plots is the median.



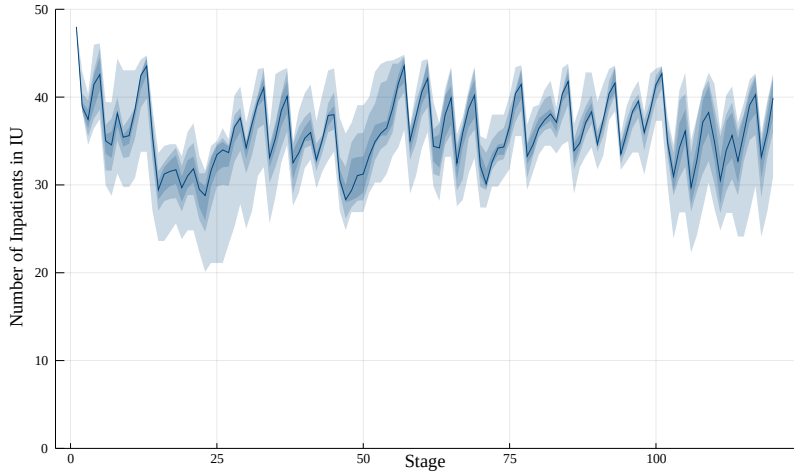
(a) Median (b) Mean

Figure 4.17: Mean and Median Costs for Current Practice vs. the MSP-PDU Policy

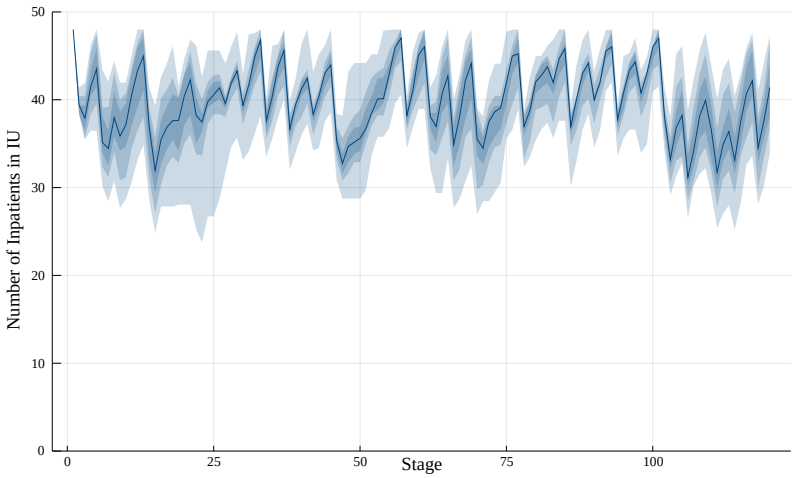
The median in Figure 4.16 is lower for the MSP-PDU policy than in the current practice. The median also has fewer spikes when a PDU exists. In addition to less frequent spikes, they are shorter with a PDU. The highest cost for the MSP-PDU policy is around $450K while it is more than $600K in the current practice. Overall, shaded area is smaller in Figure 4.16(a) vs. 4.16(b). This shows the cost is more stable and fluctuates less when the hospital assigns a separate PDU to ALC patients.

To better compare the current practice and the optimal policy, obtained from our model, only the median and mean costs are depicted in Figure 4.17. The MSP-PDU solution outperforms the current policy in terms of both the median and the mean of cost. In

some stages, the difference is a big as $7000, which is significant considering that every stage is a 6 hours time interval. The median cost of the optimal policy is smoother compared to the current practice. Further, the range of the mean and median cost for the optimal policy is smaller compared to the current practice.

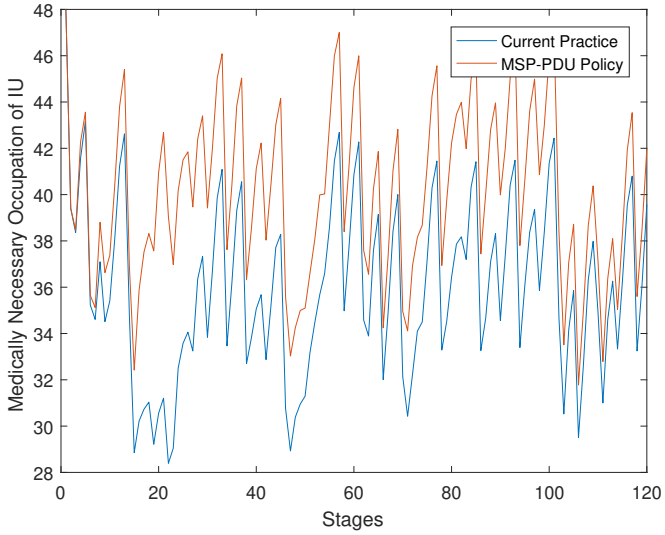


(a) Current Practice



(b) MSP-PDU Policy

Figure 4.18: Medically Needed Stays in IU



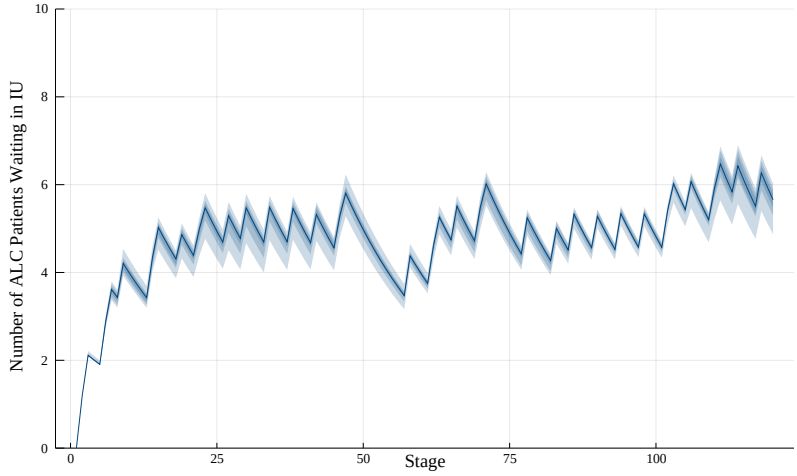
(a) Median

Figure 4.19: Median Number of Inpatients in Current Practice vs. the MSP-PDU Policy

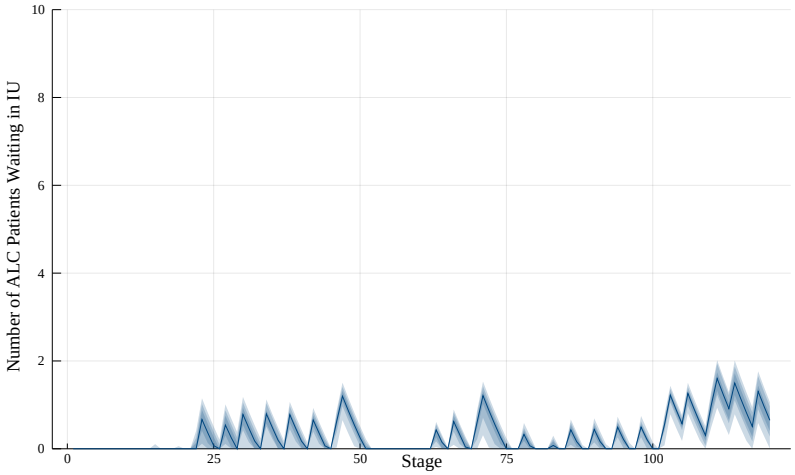
The remaining figures compare hospital congestion through a variety of performance metrics, including the number of declined bed requests, ALC population in IU, and medically necessary IU occupancy. Figure 4.18 shows the IU occupancy by inpatients who need IU beds for medical purposes. Most of the shaded region in Figure 4.18(b) is above 40, while it is between 30 and 40 in Figure 4.18(a). The minimum number of beds, used for medical purposes, is around 25 with a PDU while it is almost 20 in the current practice. The median in Figure 4.18(b) is always higher than in Figure 4.18(a). This is better shown in Figure 4.19, which shows the median of efficient IU occupancy for the two policies in the same figure. The figure for mean IU occupancy by non-ALC patients, which is in the Appendix, was very similar to the median.

Figure 4.19 shows that IU occupancy by non-ALC patients is significantly higher with a small PDU with only 6 beds. In some stages, the difference is as big as 15. Overall, Figures 4.18 and 4.19 indicate that IU beds are used more efficiently when ALC patients are transferred to a PDU. This shows that our model was able to improve access to valuable IU

beds, which is one of the main goals of this study.



(a) Current Practice

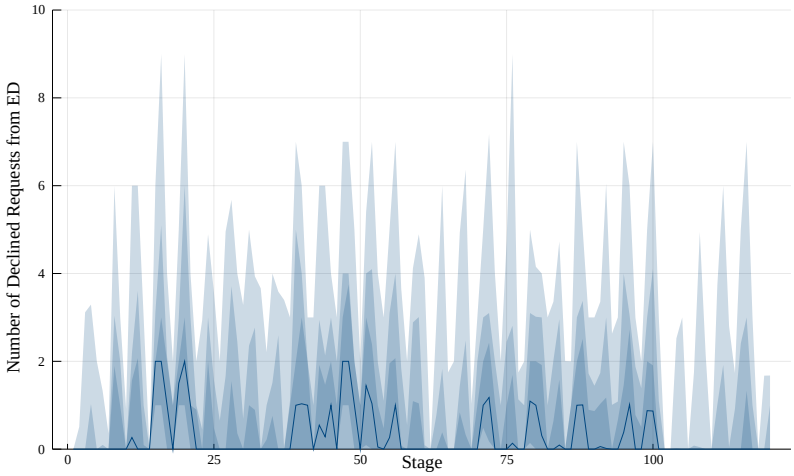


(b) MSP-PDU Policy

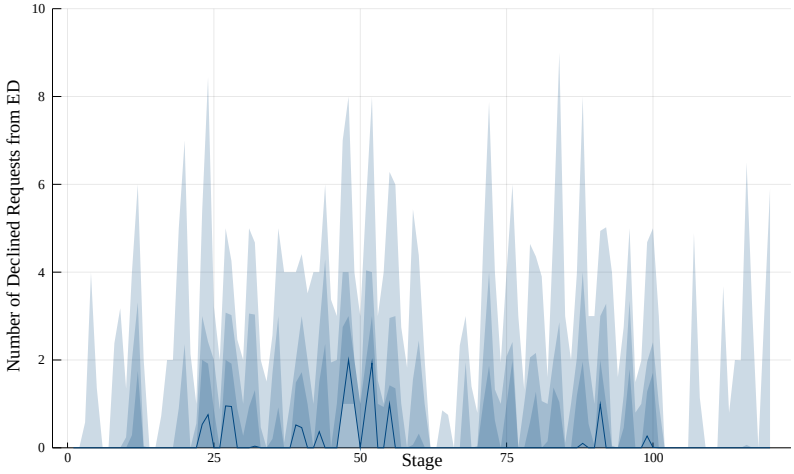
Figure 4.20: ALC Population in IU

Figure 4.20 depicts the ALC population in IU (medically unnecessary IU stays). The general trend for current practice is almost non-decreasing, which potentially can cause ED overcrowding during demand shocks for IU beds (e.g., a virus outbreak). In the current practice, between 4-6 ALC patients wait in the IU in most stages, and the worst-case is

around 7. Building a small PDU with 6 beds reduces non-medical IU stays significantly. As Figure 4.20(b) shows, the ALC population falls under 1 in most stages. Although some spikes are observable, the maximum is still under 2. As the difference between Figures 4.20(a) and 4.20(b) is evident, we do not show a separate figure for the median and mean. However, these figures are included in the Appendix.



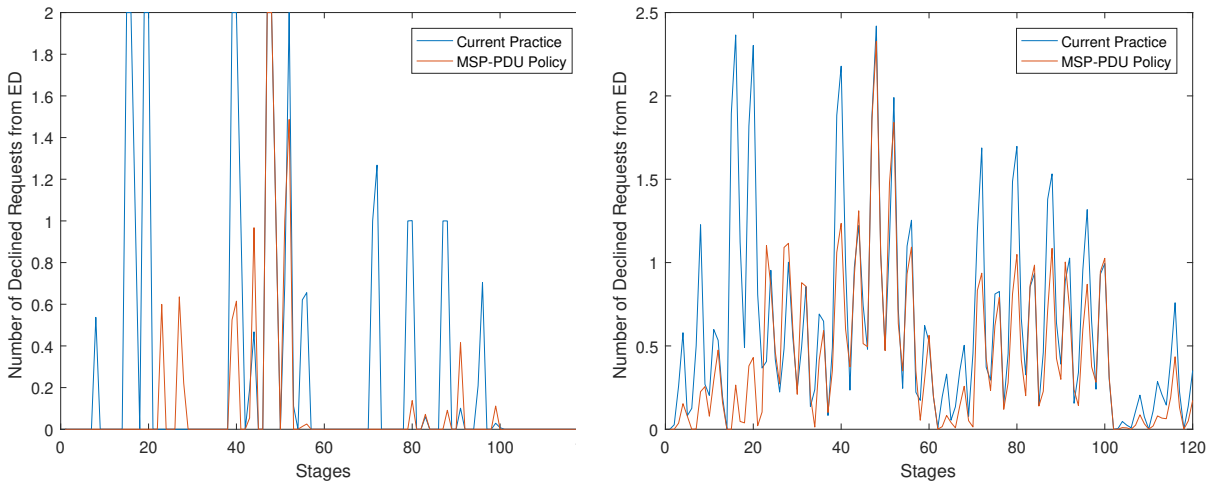
(a) Current Practice



(b) MSP-PDU Policy

Figure 4.21: ED Bed Request Rejections

Figures 4.21 display the number of declined requests for IU beds from the ED. As discussed in Section 4.4.1, the majority of bed requests in our partner hospital are made by ED. Although bed requests from other sources are more expensive to decline, they are only a few. Besides, rejecting bed requests from ED is the main contributor to hospital congestion. Figure 4.21 shows that more ED patients are declined for being admitted to IU in the current practice vs. the optimal policy. The darkest shade (25%-75% quantiles) in Figure 4.21(a) reaches higher values compared to Figure 4.21(b). In the current practice, the number of declined requests from ED goes beyond 8 in the worst cases. Although the same may happen even with a PDU, it is less frequent/likely.

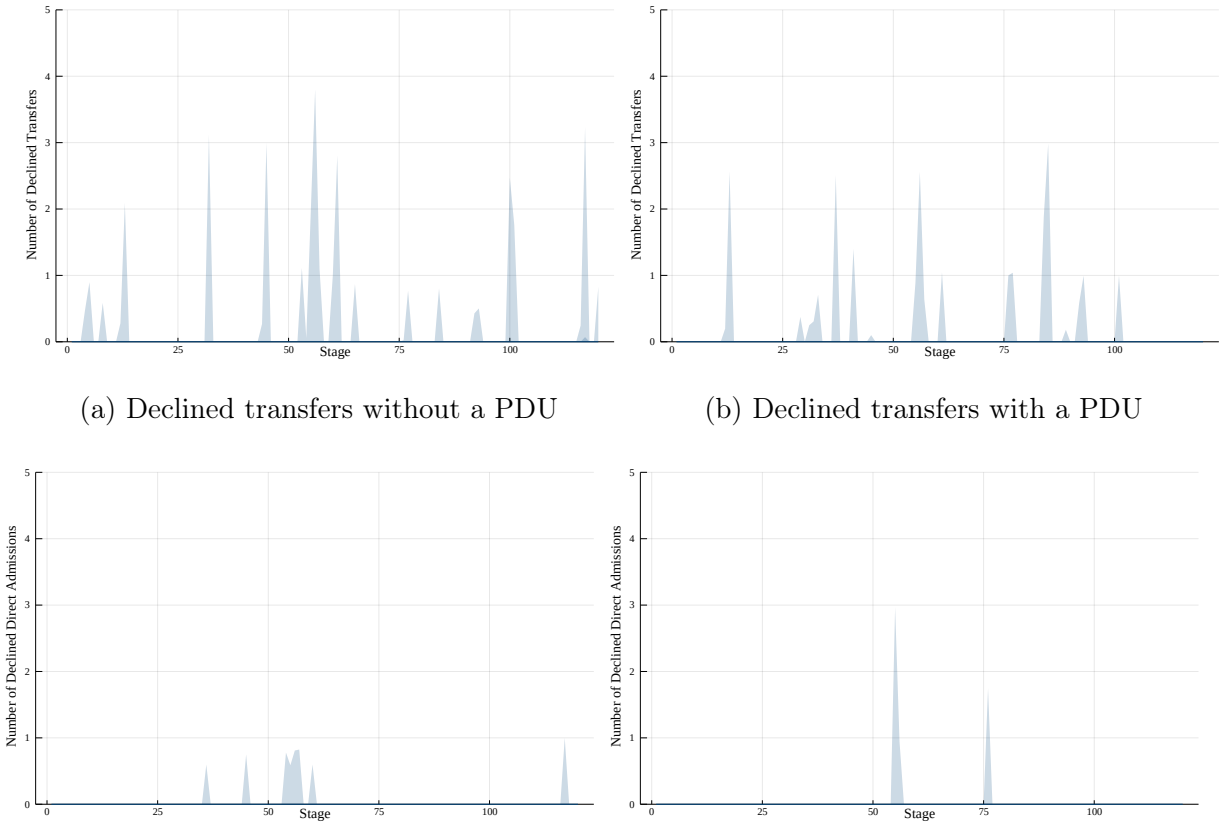


(a) Median (b) Mean

Figure 4.22: Mean and Median Number of Declined Bed Requests from ED in Current Practice vs. the MSP-PDU Policy

Further, comparing Figures 4.21(a) and 4.21(b) indicates that the median is 0 in most stages when a PDU exists while it is higher, has more spikes, and fluctuates a lot in the current practice. This is better depicted in Figure 4.22 the sub-figures of which represent the median and mean number of declined bed requests from ED in each stage. This figure

shows that, in the current practice, the spikes are significantly taller (the difference could reach 2 patients), and ED patients are more frequently (in more stages) declined compared to the optimal policy.



(a) Declined transfers without a PDU

(b) Declined transfers with a PDU

(c) Declined direct admissions without a PDU

(d) Declined direct admissions with a PDU

Figure 4.23: Transfer and Direct Admission Request Rejections

Figure 4.23 compares the number of declined transfers and direct admissions with and without a PDU. Sub-figures (a) and (b) are related to transfers, and (c) and (d) show direct admission requests. The median for the number of declined transfers and direct admissions is always 0 no matter there is a PDU in H1 or not. This is expected since G19 does not receive much transfer and direct admission requests, and those are very expensive to decline.

Thus, the MSP-PDU model always prioritizes them over bed requests from ED. However, rejecting transfer and direct admission requests still happen in less number of scenarios when a PDU exists comparing with the current practice.

Table 4.6: Current Practice vs. MSP-PDU Policy

Fixed Cost ($)	PDU Capacity	97.5% Confidence Interval of Total Cost ($K/month)	Average Medically Needed IU Occupancy (beds)
∞ (No PDU)	0	585.67 ± 4.81	35.79
8K	6.34	450.39 ± 4.50	39.9
12K	6.04	475.75 ± 4.53	39.9
16K	5.87	500.18 ± 4.59	39.9
20K	5.37	521.46 ± 4.63	39.6
24K	5.28	541.35 ± 4.54	39.5
28K	4.69	563.18 ± 4.63	39.1
32K	3.7	580.45 ± 4.67	38.2
36K	2.46	583.03 ± 4.68	37.5
40K	0	586.01 ± 4.83	35.7

Table 4.6 shows the performance of a PDU, assuming different fixed costs compared to the current practice. The table compares the 97.5% cost confidence interval and the average number of medically necessary occupied IU beds over all the stages. As expected, no PDU (current practice) is the most expensive policy and has the lowest medical usage of IU beds. The deviation among the PDU capacity for a fixed cost \$8K-\$16K is small, and the nearest integer they all round to is 6. The average medically necessary occupancy for all three fixed costs \$8K, \$12K, and \$16K is also almost the same (≈ 40). By increasing the fixed cost to 20, \$24K, and \$28K, the PDU capacity and medically needed IU usage decrease to ≈ 5 and ≈ 39,

respectively. By increasing the fixed cost beyond $28K, PDU capacity and efficient IU usage decrease more rapidly. In general, the improvement over the current practice is between 4.9%-11.6% depending on the fixed cost. Figure 4.24 shows the increase in IU bed access for

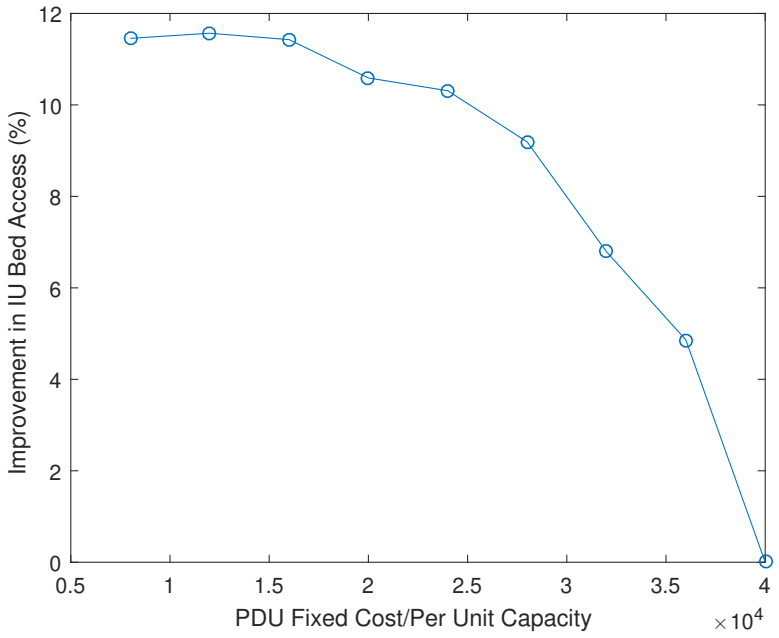


Figure 4.24: Percentage Improvement in IU Access Compared to the Current Practice

PDU with different fixed costs compared to the current practice. The results show that as long as the fixed cost is under $28K per month, the 97.5% cost confidence intervals with and without a PDU do not overlap, meaning that the differences are statistically significant.

5. CONCLUSIONS AND FUTURE RESEARCH

5.1 Conclusions

This dissertation studies different approaches to improve hospital discharge/output, which is the overlooked piece of the patient flow improvement puzzle in the literature. Ready for discharge patients are categorized based on their destination after being discharged. Besides inpatients who leave the hospital for home, many or a few might be transferred to post-acute care facilities, including long-term care hospitals and skilled nursing homes depending on patients' health complications. They usually experience several days of non-medical stay (ALC days) in the inpatient unit as hospital delays their transfer to the next level of care for a variety of reasons. Hospital's unsuccessful search to find a PACF that fits a patient's medical needs and lack of patient's decision-making capacity are among factors that contribute to ALC days. Chapter 3 in this dissertation is assigned to patients who are discharged to home. The discharge process, on the day of discharge, is studied and improved at the operational level for these patients. Chapter 4 investigates the feasibility of creating a "post-discharge-unit" (PDU) for patients who are medically ready for discharge, to a post-acute care facility, but experience transition delays.

In Chapter 3, we formulate the inpatient discharge planning (IDP) problem as a two-stage stochastic program with discharge processing time and bed request arrival time as two independent random variables. Our model optimizes the IDP problem from both the hospital and patient perspectives by minimizing upstream patient boarding and discharge lateness and by integrating patient preferences on discharge time of day. The IDP problem is solved for instances generated using real data from a large hospital in Texas. As the deterministic equivalent of the IDP two-stage model is not solvable in a reasonable time, three solution approaches, L-shaped, stochastic decomposition (SD), and the shorted expected processing time first (SEPT) heuristic are developed. Among these, SD emerges as the

only effective solution method. Overall, SD significantly improves computational time and achieves high solution quality in those cases that can be benchmarked. In current practice, a variation of the SEPT heuristic was analyzed and results show the optimization approaches are considerably better. To implement the model, a scheduling tool would be needed to interface with the model and underlying software. The nurse would use the model on a daily basis to make assignment and sequence decisions for IDP.

Chapter 4 develops a multistage stochastic program (MSP) to study the cost-effectiveness of assigning a separate unit, named post-discharge unit (PDU), to ALC patients. The model contains a multi-dimensional random variable related to the number of bed requests from different sources, including ED, ICU, transfers, and direct admissions. As the size of MSPs increases with the number of stages and outcomes per stage, solving the deterministic equivalent is not efficient. Thus we use the stochastic dual dynamic programming algorithm to solve the model. Capacity planning and cost-effectiveness of PDU is studied for a large hospital in Houston, Texas. We also assume different hospital settings by varying the ratio of transfer to PACFs. For three different hospital settings, the impact of PDU operational and fixed cost and the average number of LAC days on the cost-effectiveness of PDU is studied. We also compare the performance of a PDU in our partner hospital with the current practice in terms of cost and access to IU beds. The results show that creating a PDU in our partner hospital can save money and improve access to IU beds, although the transfer ratio to PACFs is relatively small (15%). Depending on the PDU fixed cost, it can improve access to IU beds by 4.86%-11.59% compared to the current practice, which will consequently reduce congestion in upstream units. The results also indicate that the optimal policy obtained from our model is more sensitive to PDU costs in hospitals with a larger ALC population.

5.2 Future Research

This work can be extended in several directions. The IDP problem can be formulated as a multi-stage stochastic program where every step of the discharge process, discussed in Section 3.1.1, is considered as a stage. Another extension is to formulate the IDP problem as a risk-

averse two-stage stochastic program where the risk measure minimizes the positive deviation of the discharge time from the target time. Furthermore, bed requests can be prioritized base on their urgency. For example, one can assign the highest priority to bed requests from the emergency department. Non-discharge related tasks for nurses can be incorporated into the model explicitly. Nurse availability during the day can also be included as a random variable. In addition, nurse productivity can be integrated as a decreasing parameter with respect to fatigue level and time of the day. Most importantly, steps can be taken to apply the approach in practice so that real value may be realized.

The objective function in our MSP-PDU model is cost-based. We assume that the decision-maker does not have a pre-specified target for the number of patient admissions from upstream units, more importantly, the ED. The decision maker's preference for the level of access to IU beds can be incorporated into the optimization model through a variety of methods. For example, we can assure that a target percentage of bed requests from the ED will be accepted by incorporating chance constraints in the model. Our current model assumes that requests from different sources have the same level of importance. However, declining requests from the ED contributes to hospital congestion the most. Thus, the model might prioritize bed requests from the emergency department over other sources. Further, extreme events (e.g., a virus outbreak) resulting in more extended hospital stays and higher demand for the ED and consequently IU beds, are not considered in our MSP-PDU model. A risk-averse extension of the MSP-PDU might be used to manage such extreme events.

REFERENCES

[1] S. Trzeciak and E. Rivers, "Emergency department overcrowding in the United States: an emerging threat to patient safety and public health," *Emergency Medicine Journal*, vol. 20, no. 5, pp. 402–405, 2003.

[2] American Hospital Association, "Trendwatch chartbook 2016: trends affecting hospitals and health systems." `http://www.aha.org/research/reports/tw/chartbook`, 2016.

[3] X. Shen and X. Wang, "Improving the health-care delivery process at hospital emergency services by a better use of inpatient bed information," *Electronic Commerce Research and Applications*, vol. 14, no. 1, pp. 14–22, 2015.

[4] R. W. Schafermeyer and B. R. Asplin, "Hospital and emergency department crowding in the united states," *Emergency Medicine*, vol. 15, no. 1, pp. 22–27, 2003.

[5] J. E. Helm, S. AhmadBeygi, and M. P. Van Oyen, "Design and analysis of hospital admission control for operational effectiveness," *Production and Operations Management*, vol. 20, no. 3, pp. 359–374, 2011.

[6] S. H. Kim, C. W. Chan, M. Olivares, and G. J. Escobar, "Association among ICU congestion, ICU admission decision, and patient outcomes," *Critical Care Medicine*, vol. 44, no. 10, pp. 1814–1821, 2016.

[7] C. Amy, B. Zagorski, V. Chan, D. Parsons, R. Vander Laan, and A. Colantonio, "Acute care alternate-level-of-care days due to delayed discharge for traumatic and non-traumatic brain injuries," *Healthcare Policy*, vol. 7, no. 4, p. 41, 2012.

[8] D. E. Roberts, R. G. Holloway, and B. P. George, "Post-acute care discharge delays for neurology inpatients: Opportunity to improve patient flow," *Neurology: Clinical Practice*, vol. 8, no. 4, pp. 302–310, 2018.

[9] E. J. Zhao, A. Yeluru, L. Manjunath, L. R. Zhong, H. T. Hsu, C. K. Lee, A. C. Wong, M. Abramian, H. Manella, D. Svec, *et al.*, "A long wait: barriers to discharge for long length of stay patients," *Postgraduate Medical Journal*, vol. 94, no. 1116, pp. 546–550, 2018.

[10] R. Barba, J. Marco, J. Canora, S. Plaza, S. N. Juncos, J. Hinojosa, M. M. Bailon, and A. Zapatero, "Prolonged length of stay in hospitalized internal medicine patients," *European Journal of Internal Medicine*, vol. 26, no. 10, pp. 772–775, 2015.

[11] D. Foer, K. Ornstein, T. A. Soriano, N. Kathuria, and A. Dunn, "Nonmedical factors associated with prolonged hospital length of stay in an urban homebound population," *Journal of Hospital Medicine*, vol. 7, no. 2, pp. 73–78, 2012.

[12] P. W. New, N. Andrianopoulos, P. Cameron, J. Olver, and J. U. Stoelwinder, "Reducing the length of stay for acute hospital patients needing admission into inpatient rehabilitation: a multicentre study of process barriers," *Internal Medicine Journal*, vol. 43, no. 9, pp. 1005–1011, 2013.

[13] A. P. Costa, J. W. Poss, T. Peirce, and J. P. Hirdes, "Acute care inpatients with long-term delayed-discharge: evidence from a Canadian health region," *BMC Health Services Research*, vol. 12, no. 1, p. 172, 2012.

[14] A. Swinkels and T. Mitchell, "Delayed transfer from hospital to community settings: the older person's perspective," *Health & Social Care in the Community*, vol. 17, no. 1, pp. 45–53, 2009.

[15] A. Kydd, "The patient experience of being a delayed discharge," *Journal of Nursing Management*, vol. 16, no. 2, pp. 121–126, 2008.

[16] C. Levine and K. Ramos-Callan, *The illusion of choice: why decisions about post-acute care are difficult for patients and family caregivers.* Quality Institute, United Hospital Fund, 2019.

[17] E. A. Coleman, J. D. Smith, J. C. Frank, S. J. Min, C. Parry, and A. M. Kramer, "Preparing patients and caregivers to participate in care delivered across settings: the care transitions intervention," *Journal of the American Geriatrics Society*, vol. 52, no. 11, pp. 1817–1825, 2004.

[18] K. Muthuraman and M. Lawley, "A stochastic overbooking model for outpatient clinical scheduling with no-shows," *IIE Transactions*, vol. 40, no. 9, pp. 820–837, 2008.

[19] C. Zacharias and M. Armony, "Joint panel sizing and appointment scheduling in outpatient care," *Management Science*, vol. 63, no. 11, pp. 3978–3997, 2017.

[20] N. Liu, "Optimal choice for appointment scheduling window under patient no-show behavior," *Production and Operations Management*, vol. 25, no. 1, pp. 128–142, 2016.

[21] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick, "Patient streaming as a mechanism for improving responsiveness in emergency departments," *Operations Research*, vol. 60, no. 5, pp. 1080–1097, 2012.

[22] K. Xu and C. W. Chan, "Using future information to reduce waiting times in the emergency department via diversion," *Manufacturing & Service Operations Management*, vol. 18, no. 3, pp. 314–331, 2016.

[23] D. Das, K. S. Pasupathy, C. B. Storlie, and M. Y. Sir, "Functional regression-based monitoring of quality of service in hospital emergency departments," *IISE Transactions*, vol. 51, no. 9, pp. 1012–1024, 2019.

[24] J. S. Peck, J. C. Benneyan, D. J. Nightingale, and S. A. Gaehde, "Predicting emergency department inpatient admissions to improve same-day patient flow," *Academic Emergency Medicine*, vol. 19, no. 9, pp. E1045–E1054, 2012.

[25] Y. Deng, S. Shen, and B. Denton, "Chance-constrained surgery planning under conditions of limited and ambiguous data," *INFORMS Journal on Computing*, vol. 31, no. 3, pp. 559–575, 2019.

[26] M. Bam, B. T. Denton, M. P. Van Oyen, and M. E. Cowen, "Surgery scheduling with recovery resources," *IISE Transactions*, vol. 49, no. 10, pp. 942–955, 2017.

[27] S. Mahant, R. Peterson, M. Campbell, D. L. MacGregor, and J. N. Friedman, "Reducing inappropriate hospital use on a general pediatric inpatient unit," *Pediatrics*, vol. 121, no. 5, pp. e1068–e1073, 2008.

[28] M. J. Vermeulen, J. G. Ray, C. Bell, B. Cayen, T. A. Stukel, and M. J. Schull, "Disequilibrium between admitted and discharged hospitalized patients affects emergency department length of stay," *Annals of Emergency Medicine*, vol. 54, no. 6, pp. 794–804, 2009.

[29] D. E. Holland, J. E. Pacyna, K. L. Gillard, and L. C. Carter, "Tracking discharge delays: critical first step toward mitigating process breakdowns and inefficiencies," *Journal of Nursing Care Quality*, vol. 31, no. 1, pp. 17–23, 2016.

[30] R. Srivastava, B. L. Stone, R. Patel, M. Swenson, A. Davies, C. G. Maloney, P. C. Young, and B. C. James, "Delays in discharge in a tertiary care pediatric hospital," *Journal of Hospital Medicine*, vol. 4, no. 8, pp. 481–485, 2009.

[31] R. K. Khare, E. S. Powell, G. Reinhardt, and M. Lucenti, "Adding more beds to the emergency department or reducing admitted patient boarding times: which has a more significant influence on emergency department congestion?," *Annals of Emergency Medicine*, vol. 53, no. 5, pp. 575–585, 2009.

[32] M. Kane, A. Weinacker, R. Arthofer, T. Seay-Morrison, W. Elfman, M. Ramirez, N. Ahuja, D. Pickham, J. Hereford, and M. Welton, "A multidisciplinary initiative to increase inpatient discharges before noon," *Journal of Nursing Administration*, vol. 46, no. 12, pp. 630–635, 2016.

[33] P. J. Parikh, N. Ballester, K. Ramsey, N. Kong, and N. Pook, "The n-by-t target discharge strategy for inpatient units," *Medical Decision Making*, vol. 37, no. 5, pp. 534–543, 2017.

[34] D. M. Manning, K. J. Tammel, R. N. Blegen, L. A. Larson, F. L. Steffens, D. J. Rosenman, W. C. Mundell, J. M. Naessens, R. K. Resar, and J. M. Huddleston, "In-room display of day and time patient is anticipated to leave hospital: a discharge appointment," *Journal of Hospital Medicine*, vol. 2, no. 1, pp. 13–16, 2007.

[35] E. A. Crawford, P. J. Parikh, N. Kong, and C. V. Thakar, "Analyzing discharge strategies during acute care: a discrete-event simulation study," *Medical Decision Making*, vol. 34, no. 2, pp. 231–241, 2014.

[36] G. Dobson, H. H. Lee, and E. Pinker, "A model of ICU bumping," *Operations Research*, vol. 58, no. 6, pp. 1564–1576, 2010.

[37] C. W. Chan, V. F. Farias, N. Bambos, and G. J. Escobar, "Optimizing intensive care unit discharge decisions with patient readmissions," *Operations Research*, vol. 60, no. 6, pp. 1323–1341, 2012.

[38] P. Shi, M. C. Chou, J. Dai, D. Ding, and J. Sim, "Models and insights for hospital inpatient operations: time-dependent ED boarding time," *Management Science*, vol. 62, no. 1, pp. 1–28, 2015.

[39] K. Xu and C. W. Chan, "Using future information to reduce waiting times in the emergency department via diversion," *Manufacturing & Service Operations Management*, vol. 18, no. 3, pp. 314–331, 2016.

[40] S. Enayati, O. Y. Özaltın, M. E. Mayorga, and C. Saydam, "Ambulance redeployment and dispatching under uncertainty with personnel workload limitations," *IISE Transactions*, vol. 50, no. 9, pp. 777–788, 2018.

[41] J. E. Helm, S. AhmadBeygi, and M. P. Van Oyen, "Design and analysis of hospital admission control for operational effectiveness," *Production and Operations Management*, vol. 20, no. 3, pp. 359–374, 2011.

[42] J. S. Peck, J. C. Benneyan, D. J. Nightingale, and S. A. Gaehde, "Predicting emergency department inpatient admissions to improve same-day patient flow," *Academic Emergency Medicine*, vol. 19, no. 9, pp. E1045–E1054, 2012.

[43] J. Niyirora and J. Zhuang, "Fluid approximations and control of queues in emergency departments," *European Journal of Operational Research*, vol. 261, no. 3, pp. 1110–1124, 2017.

[44] G. Allon, S. Deo, and W. Lin, "The impact of size and occupancy of hospital on the extent of ambulance diversion: theory and evidence," *Operations Research*, vol. 61, no. 3, pp. 544–562, 2013.

[45] J. K. Cochran and K. T. Roche, "A multi-class queuing network analysis methodology for improving hospital emergency department performance," *Computers & Operations Research*, vol. 36, no. 5, pp. 1497–1512, 2009.

[46] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick, "Patient streaming as a mechanism for improving responsiveness in emergency departments," *Operations Research*, vol. 60, no. 5, pp. 1080–1097, 2012.

[47] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick, "Complexity-augmented triage: a tool for improving patient safety and operational efficiency," *Manufacturing & Service Operations Management*, vol. 16, no. 3, pp. 329–345, 2014.

[48] A. P. Costa and J. P. Hirdes, "Clinical characteristics and service needs of alternate-level-of-care patients waiting for long-term care in ontario hospitals," *Healthcare Policy*, vol. 6, no. 1, p. 32, 2010.

[49] R. McCloskey, P. Jarrett, and C. Stewart, "The untold story of being designated an alternate level of care patient," *Healthcare Policy*, vol. 11, no. 1, p. 76, 2015.

[50] N. Meo, J. M. Liao, and A. Reddy, "Hospitalized after medical readiness for discharge: a multidisciplinary quality improvement initiative to identify discharge barriers in general medicine patients," *American Journal of Medical Quality*, pp. 23–28, 2019.

[51] A. Barnable, D. Welsh, E. Lundrigan, and C. Davis, "Analysis of the influencing factors associated with being designated alternate level of care," *Home Health Care Management & Practice*, vol. 27, no. 1, pp. 3–12, 2015.

[52] H. M. Bidhandi, J. Patrick, P. Noghani, and P. Varshoei, "Capacity planning for a network of community health services," *European Journal of Operational Research*, vol. 275, no. 1, pp. 266–279, 2019.

[53] F. Lin, N. Kong, and M. Lawley, "Capacity planning for publicly funded community based long-term care services," in *Community-based Operations Research*, pp. 297–315, Springer, 2012.

[54] T. Cardoso, M. D. Oliveira, A. Barbosa-Póvoa, and S. Nickel, "An integrated approach for planning a long-term care network with uncertainty, strategic policy and equity considerations," *European Journal of Operational Research*, vol. 247, no. 1, pp. 321–334, 2015.

[55] P. Intrevado, V. Verter, and L. Tremblay, "Patient-centric design of long-term care networks," *Health Care Management Science*, vol. 22, no. 2, pp. 376–390, 2019.

[56] E. N. Weiss and J. O. McClain, "Administrative days in acute care facilities: a queueing-analytic approach," *Operations Research*, vol. 35, no. 1, pp. 35–44, 1987.

[57] J. Patrick, "Access to long-term care: the true cause of hospital congestion?," *Production and Operations Management*, vol. 20, no. 3, pp. 347–358, 2011.

[58] J. Patrick, K. Nelson, and D. Lane, "A simulation model for capacity planning in community care," *Journal of Simulation*, vol. 9, no. 2, pp. 111–120, 2015.

[59] N. Koizumi, E. Kuno, and T. E. Smith, "Modeling patient flows using a queuing network with blocking," *Health Care Management Science*, vol. 8, no. 1, pp. 49–60, 2005.

[60] Y. Deng and S. Shen, "Decomposition algorithms for optimizing multi-server appointment scheduling with chance constraints," *Mathematical Programming*, vol. 157, no. 1, pp. 245–276, 2016.

[61] R. Luscombe and E. Kozan, "Dynamic resource allocation to improve emergency department efficiency in real time," *European Journal of Operational Research*, vol. 255, no. 2, pp. 593–603, 2016.

[62] B. T. Denton, A. J. Miller, H. J. Balasubramanian, and T. R. Huschka, "Optimal allocation of surgery blocks to operating rooms under uncertainty," *Operations Research*, vol. 58, no. 4-part-1, pp. 802–816, 2010.

[63] O. V. Shylo, O. A. Prokopyev, and A. J. Schaefer, "Stochastic operating room scheduling for high-volume specialties under block booking," *INFORMS Journal on Computing*, vol. 25, no. 4, pp. 682–692, 2012.

[64] E. Curcio, P. Amorim, Q. Zhang, and B. Almada-Lobo, "Adaptation and approximate strategies for solving the lot-sizing and scheduling problem under multistage demand uncertainty," *International Journal of Production Economics*, vol. 202, pp. 81–96, 2018.

[65] S. Nickel, F. Saldanha-da Gama, and H. P. Ziegler, "A multi-stage stochastic supply network design problem with financial decisions and risk management," *Omega*, vol. 40, no. 5, pp. 511–524, 2012.

[66] Y. Adulyasak, J. F. Cordeau, and R. Jans, "Benders decomposition for production routing under demand uncertainty," *Operations Research*, vol. 63, no. 4, pp. 851–867, 2015.

[67] A. B. Philpott and V. L. De Matos, "Dynamic sampling algorithms for multi-stage stochastic programs with risk aversion," *European Journal of Operational Research*, vol. 218, no. 2, pp. 470–483, 2012.

[68] S. Bruno, S. Ahmed, A. Shapiro, and A. Street, "Risk neutral and risk averse approaches to multistage renewable investment planning under uncertainty," *European Journal of Operational Research*, vol. 250, no. 3, pp. 979–989, 2016.

[69] R. Ferstl and A. Weissensteiner, "Asset-liability management under time-varying investment opportunities," *Journal of Banking & Finance*, vol. 35, no. 1, pp. 182–192, 2011.

[70] O. Y. Özaltın, O. A. Prokopyev, A. J. Schaefer, and M. S. Roberts, "Optimizing the societal benefits of the annual influenza vaccine: a stochastic programming approach," *Operations Research*, vol. 59, no. 5, pp. 1131–1143, 2011.

[71] M. Colvin and C. T. Maravelias, "A stochastic programming approach for clinical trial planning in new drug development," *Computers & Chemical Engineering*, vol. 32, no. 11, pp. 2626–2642, 2008.

[72] P. Punnakitikashem, J. M. Rosenberger, and D. B. Behan, "Stochastic programming for nurse assignment," *Computational Optimization and Applications*, vol. 40, no. 3, pp. 321–349, 2008.

[73] S. A. Erdogan and B. Denton, "Dynamic appointment scheduling of a stochastic server with uncertain demand," *INFORMS Journal on Computing*, vol. 25, no. 1, pp. 116–132, 2013.

[74] T. Schoenfelder, J. Klewer, and J. Kugler, "Determinants of patient satisfaction: a study among 39 hospitals in an in-patient setting in germany," *International Journal for Quality in Health Care*, vol. 23, no. 5, pp. 503–509, 2011.

[75] R. M. Van Slyke and R. Wets, "L-shaped linear programs with applications to optimal control and stochastic programming," *SIAM Journal on Applied Mathematics*, vol. 17, no. 4, pp. 638–663, 1969.

[76] J. R. Birge and F. Louveaux, *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[77] P. Kall and J. Mayer, "Stochastic linear programming: models, theory, and computation," *International Series in Operations Research and Management Science. Springer; New York*, 2011.

[78] J. L. Higle and S. Sen, "Stochastic decomposition: an algorithm for two-stage linear programs with recourse," *Mathematics of Operations Research*, vol. 16, no. 3, pp. 650–669, 1991.

[79] G. Infanger, "Stochastic programming," *International Series in Operations Research and Management Science. Springer; New York*, 2011.

[80] M. Pinedo, *Scheduling: theory, algorithms and systems*. Prentice Hall, Englewood Cliffs, NJ, 1995.

[81] R. Z. Farahani, A. Hassani, S. M. Mousavi, and M. B. Baygi, "A hybrid artificial bee colony for disruption in a hierarchical maximal covering location problem," *Computers & Industrial Engineering*, vol. 75, pp. 129–141, 2014.

[82] J. L. Higle and S. Sen, *Stochastic decomposition: a statistical method for large scale stochastic linear programming*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[83] M. V. Pereira and L. M. Pinto, "Multi-stage stochastic optimization applied to energy planning," *Mathematical Programming*, vol. 52, no. 1-3, pp. 359–375, 1991.

[84] G. Infanger and D. P. Morton, "Cut sharing for multistage stochastic linear programs with interstage dependency," *Mathematical Programming*, vol. 75, no. 2, pp. 241–256, 1996.

[85] J. M. Pines, R. J. Batt, J. A. Hilton, and C. Terwiesch, "The financial consequences of lost demand and reducing boarding in hospital emergency departments," *Annals of Emergency Medicine*, vol. 58, no. 4, pp. 331–340, 2011.

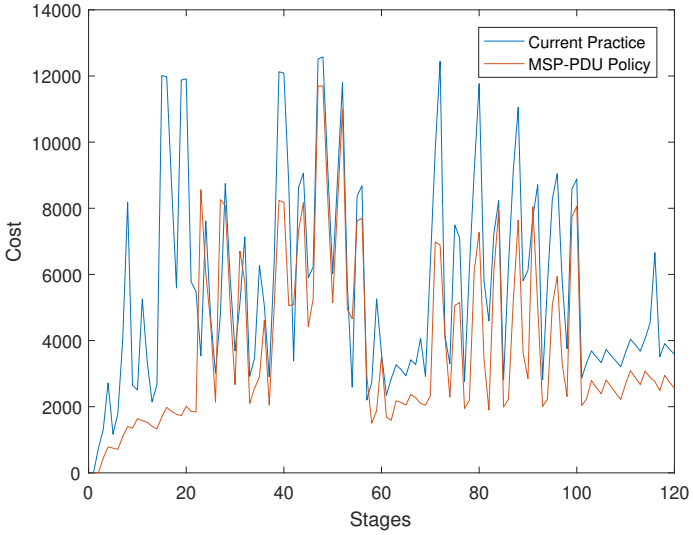
[86] C. McDonough and C. Shepard, "How can home health deliver more with less?," *The Remington Report, September/October*, pp. 3–6, 2015.

[87] C. J. Poulos, K. Eagar, and R. G. Poulos, "Managing the interface between acute care and rehabilitation-can utilisation review assist?," *Australian Health Review*, vol. 31, no. 5, pp. 129–139, 2007.

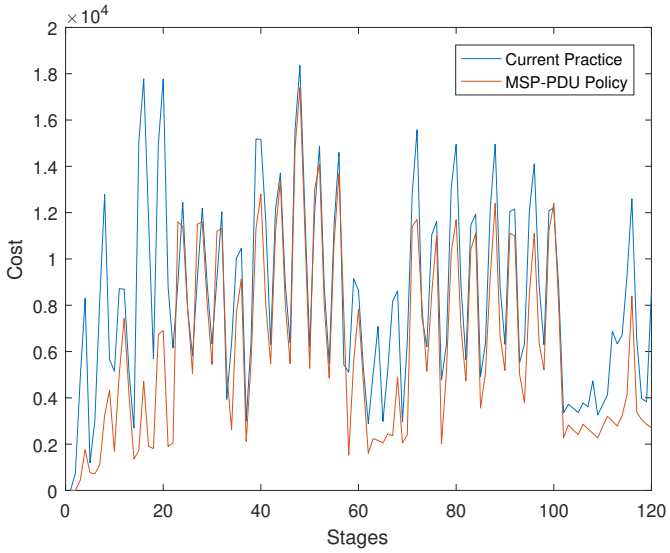
[88] O. Dowson and L. Kapelevich, "SDDP.jl: a Julia package for stochastic dual dynamic programming," *Optimization Online*, 2017.

APPENDIX A

COMPUTATIONAL RESULTS FIGURES



(a) 75% Quantile



(b) 90% Quantile

Figure A.1: Costs for Current Practice vs. the MSP-PDU Policy

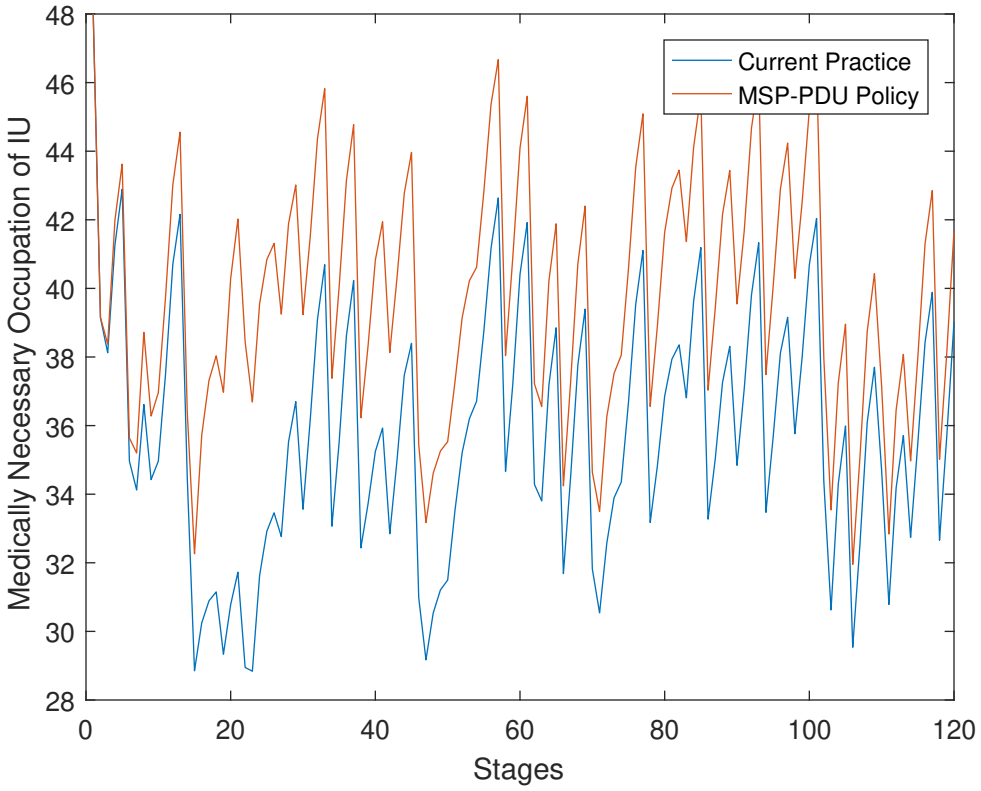
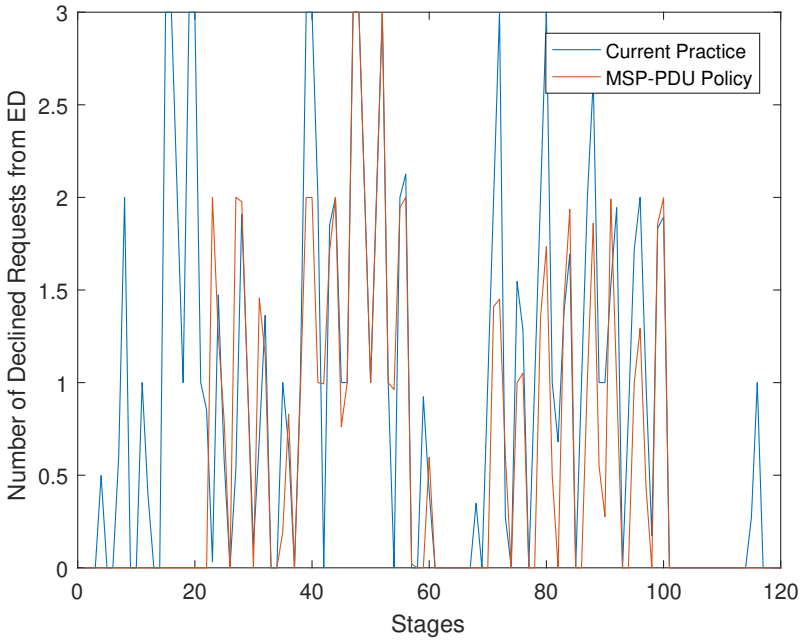
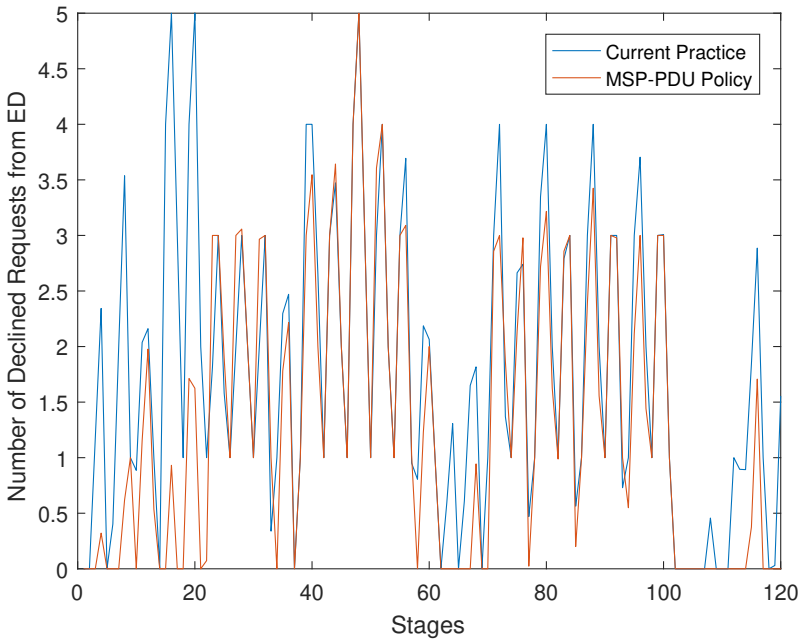


Figure A.2: Mean IU Medical Stays for Current Practice vs. the MSP-PDU Policy

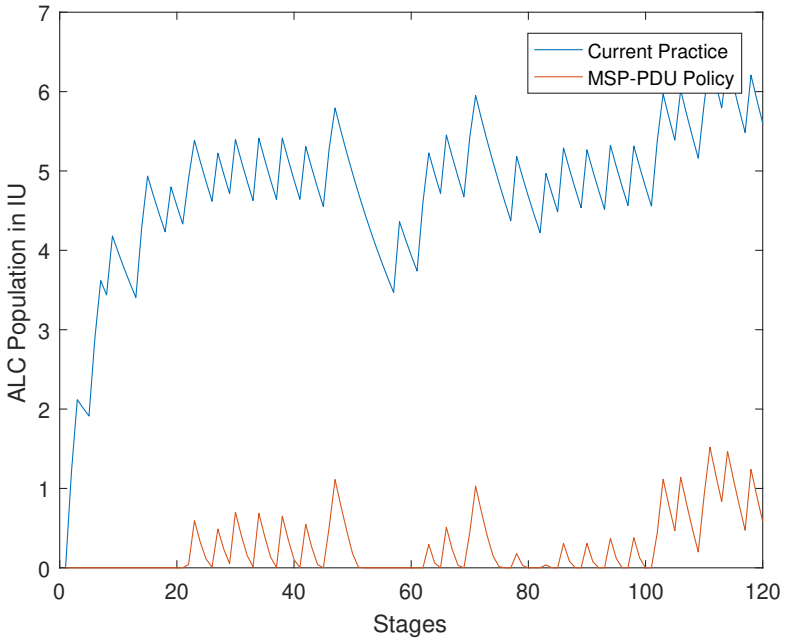


(a) 75% Quantile

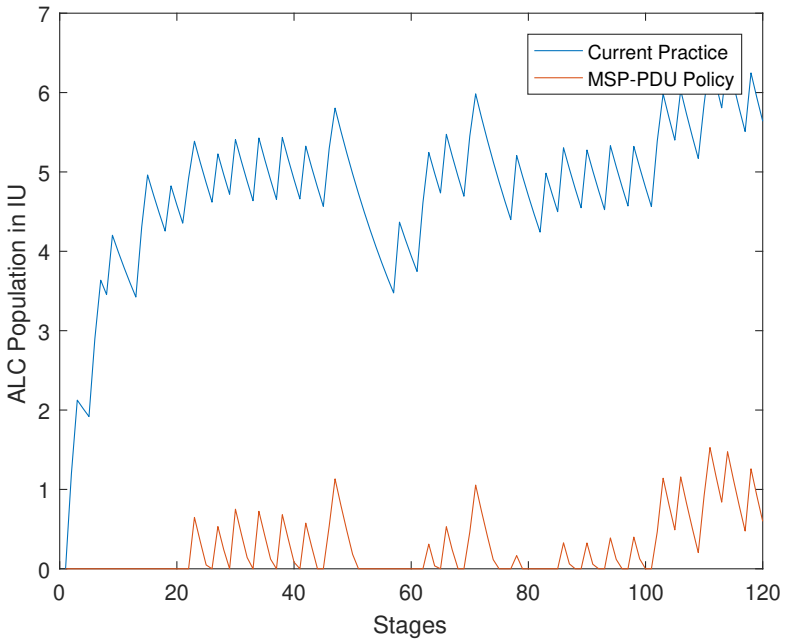


(b) 90% Quantile

Figure A.3: Declined Requests from ED for Current Practice vs. the MSP-PDU Policy



(a) Mean



(b) Median

Figure A.4: ALC Population in Current Practice vs. the MSP-PDU Policy