OPERATIONAL DECISION MAKING ACROSS PATIENT CARE CYCLE: FROM CAPACITY PLANNING TO CARE MANAGEMENT

By

Hyo Kyung Lee

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Industrial and Systems Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2019

Date of final oral examination: 05/31/2019

The dissertation is approved by the following members of the Final Oral Committee:
Jingshan Li, Professor, Department of Industrial and Systems Engineering
Laura Albert, Associate Professor, Department of Industrial and Systems Engineering
Douglas Wiegmann, Associate Professor, Department of Industrial and Systems Engineering
Gabriel Zayas-Caban, Assistant Professor, Department of Industrial and Systems Engineering
Marlon Mundt, Associate Professor, Department of Family Medicine and Community Health

To my parents

Acknowledgement

I am and always will be indebted to my PhD advisor, Professor Jingshan Li. I would like to express my sincere gratitude for his continuous guidance, patience, and encouragement. Without his thoughtful encouragement and careful supervision, this dissertaion would not have been possible. I am blessed to have him as my advisor. I also heartfully thank my committee members, Professor Laura Albert, Professor Douglas Wiegmann, Professor Gabriel Zayas-Caban, and Professor Marlon Mundt for their invaluable contributions to the direction and richness of this dissertation. I would like to thank Dr. Albert Musa and Dr. Philip Bain from Dean Health Systems for the numerous collaboration opportunities throughout my graduate study. I would also like to thank Dr. Yue Dong and Dr. Brian Pickering from Mayo Clinic for the summer research experience.

I must also thank my friends and fellow graduate students. I especially appreciate my academic siblings Xiang Zhong, Cong Zhao, Feng Ju, Sujee Lee, and Wenjun Zhu who are also great friends to me. I would like to thank my friends Hyo Jeong Kang, Ja Young Lee, and Mengyue Wang for their support along this journey as well. More than anything, I thank Seong Jae Hwang for standing by me through the ups and downs of life and the PhD program. Thank you for making my days in Madison a warm and happy memory.

Lastly, I have the greatest gratitude and love for my parents and my sister. They supported all my aspirations in life which always gave me strength facing challenges in my life. I love you with all my heart and thank you for standing by me. This dissertation is lovingly dedicated to them.

Abstract

In recent decades, healthcare has become increasingly expensive, creating pressure on care providers to deliver quality care while reducing costs. Consuming almost 20% of the gross domestic products (GDP), healthcare is one of the largest industry sectors in the United States (US). However, such high expenses may not always result in an adequate healthcare service quality where long waits, limited access and resource overloads are commonly observed.

Now more than ever, efforts to deliver care more efficiently and effectively are being pursued throughout the US. Moreover, the US healthcare system is facing incredible challenges as healthcare is shifting from the traditional, volume-driven, fee-for-service model towards value-based payment and care delivery model. Such value initiatives are becoming increasingly prevalent and important as patient costs continue to rise and access to affordable care is threatened.

Given such widespread movements toward value-based care models, all provider segments have a more growing role to play in quality improvement and enhancing their care delivery at all points along the patient care continuum. That said, the healthcare industry can no longer focus on acute care providers and hospitals only; post-acute, sub-acute, and non-acute providers play an increasingly important role as patient outcomes are being tied to readmissions and value-based payments, increasing the importance of all provider roles across the care delivery cycle.

This dissertation is dedicated to improving the efficiency and quality of the healthcare system across the care delivery cycle: from prevention to diagnosis to treatment and to

home care. Specifically, analytical tools and models to support systematic and evidence-based decision making are introduced for each stage within the cycle.

Starting from the prevention stage, we introduce a Markov chain based modeling framework to assess the impact of implementing a new service model in primary care clinics. An application study at Dean East Clinic of SSM Health is presented and managerial insights from the model regarding the impact of various workload allocation policies are discussed as well. In the following chapter, focusing on the diagnosis stage, a system-theoretic method is introduced to analyze the diagnosis-to-treatment process for lung cancer patients. As the process commonly involves frequent and potentially harmful delays, speeding up the timeliness without sacrificing the care quality is critical to improve patient outcome as well as satisfaction. To do so, we decompose the complex care delivery process to evaluate the system performance and derive indicator measures that can be used to identify the bottleneck waiting steps. Moreover, the complete distribution of the total waiting time is formulated to estimate the probability to receive the surgery within a desired or given time period. Finally, the applicability of the proposed method is illustrated via a case study at Baptist Memorial Hospital. For the treatment stage, the next chapter investigates the delays or blockings that occur during the intra hospital patient flow between different departments. Specifically, a finite capacity queueing network model based iterative procedure is formulated to evaluate the transition delay times, average bed occupancy rates, and probabilities of full occupancy. Finally, we investigate system properties to provide managerial guidance to improve patient transitions and reduce delays. To complete the patient care cycle analysis, the subsequent chapter focuses on the last stage of the care cycle: postdischarge or home care phase. As there are considerable variations in the postdischarge care process

for total joint replacement (TJR) patients, we formulate the TJR postdischarge intervention process as a finite-horizon discrete-time Markov decision process. Specifically, we dynamically model the post-TJR intervention process by directly incorporating the readmission risk and penalty, and considering the varying effectiveness of interventions depending on where the patient is located at. The applicability of the model is illustrated through a case study at St. Mary's Hospital where the derived optimal policy provides guidance to healthcare professionals in determining the optimal timing and target group of interventions.

In summary, the work developed in this dissertation provides quantitative tools to support operational decision making across the patient care cycle, and ultimately contributes to delivering safe and patient-centered care in a coordinated and seamless system.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Challenges in Healthcare Delivery

In recent decades, healthcare expenditures in the United States (US) have been growing rapidly, creating pressure on healthcare providers to cut costs while maintaining or improving care quality [1]. As one of the largest industry sectors, healthcare consumes almost 20% of the gross domestic products (GDP). However, such high expenses may not always result in an adequate healthcare service quality. Although the quality of medical care is improving for most types of illness, the attention to the processes that transform resources into healthcare services has not kept pace [2], resulting in long waiting times and delays, difficulty in care access, and healthcare provider burnout [3]. As such, US healthcare costs remain high at $3.2 trillion spent annually, of which an estimated 30% is related to waste, inefficiencies, and excessive prices [4].

Moreover, the healthcare industry is facing incredible challenges as healthcare is shifting from traditional, volume-driven, fee-for-service to value-based payment and care delivery model. One major driver is the Medicare Quality Payment Program (QPP), authorized under the 2015 MACRA legislation, which adjusts how doctors are reimbursed for services based on quality and cost [5]. The QPP program went into effect in 2017 and since then, has been encouraging providers for better care delivery. Throughout the US, value initiatives are becoming increasingly prevalent and important as patient costs

continue to rise and access to affordable care is threatened. Now more than ever, efforts to deliver care more efficiently and effectively are being pursued.

Due to such widespread movements toward value-based care models, all provider segments have a growing role to play in quality improvement and enhancing their care delivery at all points along the care continuum. The healthcare industry can no longer focus on acute care providers and hospitals only; providers across the care delivery cycle will be playing more equal roles in patient outcomes and value strategies. For instance, post-acute, sub-acute, and non-acute providers are becoming increasingly important as patient outcomes are being tied to readmissions and value-based payments, increasing the importance of all provider roles across care settings.

1.2 Care Delivery Cycle Across the Continuum

From a macro perspective, a patient's care life cycle refers to care provided from birth to end of life since an individual constantly makes decisions and engages in activities that affect his or her health [6]. A patient care cycle consists of multiple activities (from screening to diagnosis to treatment and to postdischarge care), phases (from pre-operative, operative and to post-operative), and facilities (from primary care clinic to specialty clinic to hospital and to home care). In this disseration, the care delivery cycle is defined in terms of the activities or stages patients go through: starting from a healthy state, a patient has regular preventive visits to primary care clinics to maintain good health status (and also treat minor illness that is curable through such visits). Once the patient develops an illness that requires more medical attention, he/she goes through multiple diagnosis procedures to confirm the disease stage, then receives appropriate treatment either in hospital or outpatient clinic settings. After discharge, home care

visits are conducted to ensure a safe and fast recovery back to healthy state. This complete cycle of activities encompassing the healthcare delivery system is depicted in Figure 1.1.

Figure 1.1: Care Delivery Cycle

Broadly speaking, the goal of such care delivery cycle is simple: maximizing the proportion of time a patient spends in "healthy living" stage, which in turn is equivalent to minimizing the proportions in all other stages. Hence, there exist scopes for operational improvements by reducing the inefficiencies or delays experienced by patients at each stage (prevention, diagnosis, treatment, home care) through effective capacity planning, resource allocation and care management. Improving every stage of the care cycle contributes to delivering safe and patient-centered care in a coordinated and seamless system.

This disseration is dedicated to improving the efficiency and quality of the healthcare system across the care delivery cycle. Specifically, analytical tools and models to support systematic and evidence-based decision making are introduced througout the chapters. Such tools would ideally improve health system by: (1) improving access to care; (2) reducing unnecessary waiting time; and (3) reducing cost by optimizing resource allocation. Each chapter is devoted to each stage within the care delivery cycle where a set of techniques and strategies that can be used by clinicians and administrators to improve efficiency will be introduced. The corresponding stage within the patient care cycle,

scope, and approach for each chapter are presented in Table 1.1.

Table 1.1: Chapter scopes, approaches, and corresponding stages in care cycle

Chapter	Stage	Scope	Approach
3	Prevention	Access to primary care	Markov chain
4	Diagnosis	Timeliness in diagnosis-to-treatment	Systems modeling
5	Treatment	Delays in inpatient transitions	Queueing theory
6	Home Care	Postoperative care management	Markov decision process

1.3 Prevention: Improving Access to Primary Care

Primary care is the backbone of the US healthcare system. However, it is facing tremendous challenges due to difficulties in providing timely access to patients and balancing insurmountable workloads of physicians. The demand for primary care services has increased substantially due to population growth and aging as well as the expanded healthcare insurance coverage. Recent studies show that 62 million people nationwide have no or inadequate access to primary care [7], and a significant part of physicians' tasks are focused on non-direct care activities such as clerical and administrative work [8]. As such, improving patient access and/or reducing physician workload in primary care clinics are in critical need.

To achieve this, substantial amount of efforts have been devoted to redesigning the primary care systems mainly through remodeling the provider work flow and patient flow in primary care clinics [9, 10]. Specifically, as one of such efforts, joint visit has been introduced to improve operational efficiency and reduce provider workload. Joint

visit refers to the process where the support staff assists the provider to document notes into the electronic medical record (EMR) system in real-time while the provider focuses on the patient. Hence, by shifting some portion of non-direct care workload from the provider to the support staff, joint visit system is expected to reduce provider's burden and improve care efficiency and quality as well. However, except for some pilot studies, there is no method available to rigorously investigate the impact of such service models. It is still not clear how to implement the joint visit system in practice and how to redistribute the workload among care providers, and what impact it might have on patient flow as well as provider and staff utilizations. Therefore, there is a need to develop quantitative models of joint visits in primary care clinics to evaluate the system performance and investigate the impact of various workload allocation policies.

1.4 Diagnosis: Timeliness in Lung Cancer Diagnosis-to-treatment

Lung cancer is the number one cause of cancer deaths in both men and women in the US. In 2016, there have been approximately 224,390 new cases in the US, representing about 13% of all cancer diagnoses, and 158,080 estimated deaths, accounting for approximately 27% of all cancer deaths [11]. About 402,326 Americans living today have been diagnosed with lung cancer at some point in their lives. Lung cancer is also the most common cancer worldwide, with 1.6 million new cases and accounting for approximately 1.4 million deaths annually [12]. In addition, lung cancer has a much lower five-year survival rate (17.8%) than many other leading cancer sites such as colon (65.4%), breast (90.5%) and prostate (99.6%) cancers. In fact, it is identified that more than half of the people with lung cancer die within one year of diagnosis. Part of the problem is that only 16 percent

of lung cancer cases are diagnosed at an early stage. As the lung cancer detection, diagnosis, and selection of the most appropriate treatment can be difficult, it is common to observe frequent and potentially harmful delays, as well as well-documented quality of care heterogeneity.

The diagnosis-to-treatment process for lung cancer patients is a complex, long, and stage dependent process involving multiple diagnoses, staging, treatment selections, tests, procedures, as well as different types of specialists. A lung cancer patient's journey typically starts with an abnormal chest X-ray and/or a CT-scan which is typically followed by a diagnostic biopsy and noninvasive or invasive staging tests. Treatment is stage-dependent with options including surgery, chemotherapy, radiation therapy, or a combination of these modalities. However, even though there exists a standard diagnosis procedure, its execution depends on a number of issues and has a high variation. For instance, some diagnostic tests may be skipped, some may be taken in a reversed sequence, and some tests may need to be taken multiple times.

Although most tests and procedures can be done within minutes or hours, inevitable but substantial waiting times can occur between the tests and procedures. In fact, days or even months of delays are not uncommon. Such delays for a potentially life-threatening illness not only lead to unpleasant experience to both the patients and care providers, but also may be linked to adverse survival rate. Therefore, without sacrificing the care quality, speeding up the diagnosis-to-treatment process is critically important to improve patient outcome as well as satisfaction. Specifically, identifying the most impeding waiting times (i.e., bottlenecks) can facilitate quality improvement by directing attention to specific opportunities to improve care delivery in the most efficient manner. To identify such bottlenecks, a rigorously quantitative approach is needed since the numerous alternative pathways to care implies that neither the waiting time with

the longest duration nor the waiting time that most patients experience may necessarily be the most impeding one. Hence, developing a rigorous quantitative model of the lung cancer diagnosis-to-treatment process is critically needed to identify the constraints to facilitate quality improvement.

1.5 Treatment: Reducing Delays in Inpatient Transitions

Healthcare systems in general, and hospitals in particular, require an extensive amount of resources and at the same time, suffer from a significant number of inefficiencies such as long waits, delays, cancellations, and resource overloads. The traditional approach to dealing with such inefficiencies has simply been adding more resources: building more beds, adding more staffs, etc. [13]. However, this approach has become increasingly impractical as hospitals are facing significant expenditure constraints while at the same time trying to improve care quality [14]. Due to the lack of space or the funds to expand or add resources, more and more hospitals are being forced to look at improving their patient flow or workflow by studying constraints and limitations in their process that artificially add to the problem.

Patient flow in hospitals is of particular interest to both researchers and practitioners since improving it can have a significant impact on quality of care as well as on patient satisfaction. Within a hospital, patient flow refers to the movement of patients through different units (i.e., patient transitions) such as transferring between emergency department (ED), intensive care unit (ICU), general wards, surgery suites, testing laboratories, etc. Such transitions are an important aspect of patient care which plays a critical role in providing the continuity of medical care and thus is directly linked to patient outcomes.

However, modeling patient transitions is considered to be very complex because of the different pathways patients may take and the inherent uncertainty and variability of healthcare processes. Moreover, due to limited resource, queues may be formed during such transitions.

In fact, significant delays could happen during the transitions which can lead to negative patient outcomes. For example, during the transitions from ED to ICU, barriers due to overcrowding and communication issues have led to three to five hours long waiting time that could result in delayed treatment, poor outcomes and increased mortality [15]. It is identified that only 31.2% of patients can be immediately transferred to ICU where survival rates can be decreased by 30% due to such delays [16]. Similarly, in ED boarding process, up to 25% of patients admitted to the wards are delayed for more than four hours at ED, which posts high risks in compromised patient care and safety [17]. The ICU bumping (discharging ICU patients earlier than they should to accommodate the arrivals of new critical patients) could sacrifice the bumped patients and increase their readmission probability back to ICU [18]. Therefore, studying patient transitions to reduce delays is essential in improving hospital operations and patient outcomes. Specifically, to understand the complex interactions among different units, development of a quantitative model to systematically study transitions across all interconnected units is needed.

1.6 Home Care: Optimal TJR Postoperative Care Management

In 2011, 3.3 million hospital readmissions (when a patient is admitted to a hospital within a specified time period after being discharged from an earlier hospitalization) occurred

with an associated cost of $41.3 billion. To address the problem, the Centers for Medicare and Medicaid (CMS) launched the Hospital Readmission Reduction Program in 2012 to penalize hospitals for high readmission rates and incentivize providers to improve care quality. Specifically, CMS tracks readmissions for Medicare patients admitted initially for six targeted conditions: heart attack, heart failure, pneumonia, chronic obstructive pulmonary disease (COPD), elective hip or knee replacement (total joint replacement (TJR)), and coronary artery bypass graft (CABG). Among the six conditions, TJR comprises the largest procedural expenditure in the Medicare budget [19], hence is an appropriate target for quality improvement and cost containment

TJR refers to a surgical procedure in which parts of an arthritic or damaged joint are replaced with a prosthesis to increase mobility and decrease discomfort for individuals whose pain prior to surgery could not be managed with medications or physical therapy [20]. TJR procedures are performed for more than a million cases each year where the numbers are expected to continue to increase dramatically as the US population ages; TJR has been projected to increase to 572,000 total hip arthroplasty (THA) cases and 3,480,000 total knee arthroplasty (TKA) cases by 2030 [21]. Successful outcomes from TJR surgery depend not only on the surgery quality, but also on the patient's post-operative behaviors such as leg exercises, ambulation, position change, deep breathing and coughing [22]. Inadequate performance of such postoperative behaviors can lead to complications such as deep vein thrombosis, pulmonary embolus, and pneumonia [23], which may result in unplanned hospital readmissions. In fact, up to 11% of patients are readmitted to the hospital soon after the procedure [24] and the cost of such readmissions is estimated to be $17.4 billion. Of those, CMS estimates 20% of readmissions are preventable, representing an estimated annual savings greater than $2 billion [25].

The key aspects in the postoperative care of TJR patients are functional mobility

restoration and adequate pain management [26]. To do so, most patients receive postoperative physical therapy (PT) and/or some form of rehabilitative services ranging from inpatient stays to home care services and outpatient rehabilitation, or a combination of both [27, 28]. However, there are considerable variations in postoperative care program delivery and duration; that is, there exists a wide variation in the setting, timing, amount and approaches of intervention [29]. Due to such variation, there is no agreement on which interventions given in which timeframes lead to optimal patient outcomes. As concluded by a National Institutes of Health (NIH) conference, rehabilitation services are still the most understudied aspect of the peri-operative management of TKA patients [30].

Given the continued increase in the number of joint replacement surgeries, further investigation is needed to justify the postoperative processes pertaining to TJR surgeries [28]. The costs of providing postsurgical interventions can be significant but are frequently undiscussed when considering the expenses related to TJR surgery [31]. It is estimated that the national average total postdischarge expenditures exceed $3.4 billion [28], thus attention should be directed at defining more appropriate interventions to minimize postdischarge costs while maximizing functional outcomes. As postdischarge costs (including the cost of readmissions) are one of the largest contributors of the total cost of care for TJR patients, optimized strategies to minimize postoperative care costs without compromising patient clinical outcomes are in need [32].

1.7 Organization of the Document

The rest of this document is organized as follows. Chapter 2 reviews the related literature for each stage within the patient care cycle: prevention, diagnosis, treatment and home

care. Chapter 3 proposes a Markov chain based modeling framework to assess the impact of implementing a new service model in primary care clinics and provides guidelines in adopting the appropriate form of service model. In Chapter 4, to accelerate the diagnosis-to-treatment process for lung cancer patients, a systems modeling approach is introduced to evaluate the performance and identify the bottlenecks, i.e., the most impeding waiting steps. Moreover, the complete distribution of the total waiting time is derived to capture the variability of the process. Chapter 5 focuses on the delays or blockings that can occur during the inpatient transitions for hospitalized patients. A finite capacity queuing network-based iterative procedure is developed to model the intra hospital patient flow between different units in a hospital. Such a model can serve as a quantitative decision making tool for managing bed capacity and identifying optimal allocation of hospital resources. In Chapter 6, the TJR postoperative intervention care process is formulated as a finite-horizon discrete-time Markov decision process. Utilizing the model, we investigate the decision problem faced by healthcare professionals: when should we provide interventions to which group of patients to minimize the total expense. Optimal policies derived from the model can inform the development of evidence-based clinical practice guidelines. Finally, the summary and possible extensions are presented in Chapter 7. All proofs and derivations can be found in the Appendices.

Chapter 2

Literature Review

The goal of this disseration is to develop analytical models to improve efficiencies and support decision making in healthcare delivery systems across the patient care cycle. Specifically, the focus is in seeking operational improvements at each stage of the cycle: prevention, diagnosis, treatment and home care. In this chapter, each section provides literature review for each care stage: Section 2.1 highlights the research efforts on re-designing primary care. Section 2.2 reviews the studies on delays and waiting times during the diagnosis-to-treatment process for lung cancer patients. Prevailing literature on inpatient transitions is summarized in Section 2.3. Finally, Section 2.4 reviews the existing research on postoperative care processes for joint replacement patients.

2.1 Improving Access to Primary Care

Primary care redesign has been a center topic in healthcare operations improvement movement, hence attracting a substantial amount of research efforts (see reviews in [9] and [10]). A vast majority of prevailing literature are qualitative or empirical studies that encompass team work, staffing, EMR, information systems, medical homes, payment systems, and scheduling where examples of these studies can be found in [33, 34, 35, 36, 37, 38, 39, 40, 41].

Quantitative models focusing on primary care clinics can be roughly divided into

two categories: simulation and analytical methods. Discrete-event simulations have been extensively used in studying primary care clinics where appointment scheduling, staff allocation, patient arrivals, etc., have been the major issues addressed. Examples of these studies can be found in [10, 42, 43, 44, 45].

As alternatives to simulations, analytical methods such as queueing and Markov chain models can provide rigorous analysis based on mathematical modeling. An example of queueing models application for patient cycle time evaluation and capacity design in urgent care setting is described in [46]. Additional reviews of queueing models in healthcare delivery systems are presented in [47] and [48]. Markov chain has also been extensively used to model the work flow in healthcare systems where some common areas of interest are specialty and testing clinics such as gastroenterology [49], computed tomography [50] and mammography testing [51]. However, the use of analytical models has been much more limited in studying primary care operations. Of the limited studies, the impact of implementing the electronic visit system (i.e., communication of physician and patient through a secure portal) in primary care clinics is studied in [52] and the care delivery activities within a patient room in primary care clinics are analyzed in [53].

As a way to improve primary care operation efficiency and reduce burdens of physicians, joint visit systems have been utilized in many clinics [54]. However, no quantitative study has been carried out to evaluate the different joint visit service models and compare their performances. It can be expected that analytical models of such service models can assist in the overall understanding of the joint visit system and provide guidelines in adopting the appropriate form for implementation.

2.2 Timeliness in Lung Cancer Diagnosis-to-treatment Process

Numerous studies on delays and waiting times during the lung cancer diagnosis-to-treatment process have been carried out from a clinical point of view. For example, a literature review of 1121 papers is presented in [55] to investigate whether waiting times and delays have any bearing on prognosis and treatment. It shows that evidence is needed for the prognostic impact of delays when evaluating the efforts of early detection and reducing waiting times. Through systematically reviewing the studies describing timeliness of care, and examining the associations between timeliness and clinical outcomes in patients with lung cancer, paper [56] concludes that the times to diagnosis and treatment of lung cancer are often longer than recommended. The waiting times of 29 lung cancer patients are analyzed in [57], and it discovers that limited access to specialists is the most frequent reason for the poor performance in treating lung cancer. It is shown that the delay between diagnostic and planning CT scans could result in an increase in the cross-sectional tumor size up to 373%. Paper [58] measures delays in diagnosis process for 132 patients with lung cancer at Turku University Hospital in Finland during 2001, and analyzes the causes of delays and the relationship between delay times and survivals. It suggests a multidisciplinary team approach and rapid access to carefully planned investigations to shorten the diagnostic and treatment delay times. The Time to Treat Program, which is designed for patients with clinical or radiographic suspicion of lung cancer, is introduced in [59]. The results show that the program is effective in shortening the time from suspicion of lung cancer to diagnosis and it reduces

time intervals at each step in the process. Moreover, it concludes that earlier diagnosis of lung cancer may allow increased treatment options for patients and may improve outcomes.

While there exists abundant clinical studies, quantitative modeling approach has been less adopted in studying the diagnosis-to-treatment procedure. Paper [60] introduces the process stages in lung cancer diagnosis and presents a conceptual framework to model the process. However, explanation of the detailed analysis, in particular, the evaluation of variability is not discussed. Hence, despite the importance of reducing delays in the lung cancer diagnosis-to-treatment process, there exists a huge gap in prevailing literature on generalizing how this timeliness can be achieved. As quantitative modeling allows generalization of the process, sources of delays could be rigorously investigated by disseminating the model, which in turn could facilitate quality improvement by directing attention to specific opportunities.

2.3 Reducing Delays in Inpatient Transitions

Extensive research efforts have been devoted to studying patient transitions from clinical or pilot study perspectives. For example, 43 transition studies are examined in [61] by reviewing the literature from 1982 to 2003. It is argued that theoretical framework, reliable instruments, and adequate controls are needed for transition research. A review of 21 clinical trials on transitional care for chronically ill adults is presented in [62] and quality measures to assess care transitions for hospitalized children are investigated in [63].

For transitions between ED and ICU within hospitals, the potential impact of ED overcrowding on the critically ill ED patient is reviewed in [64], which suggests that

insufficient beds is one of the main reasons for ED crowding. Similarly, the relationship between ED boarding to ICU for critically ill patients and the associated outcomes are investigated in [65], and boarding to neuro-ICU for severe stroke patients is studied in [66]. As ICU delays have substantial impact on patient safety, an ICU admission study is presented in [16] which shows that among 401 patients, 276 have been delayed which contributes to 30% of mortality risk. In addition, the ICU death rate increases 1.5% for each additional waiting hour.

For transitions to hospital ward, the vulnerabilities in ED to ward transfers for internal medicine patients are investigated in [67] by surveying the ED house staff, ED physician assistants, internal medicine house staff and hospitalists at an urban academic medical center. In [68], the effect of ICU capacity on the process and care outcome for sudden clinical deteriorating patients in the hospital wards is evaluated.

While there exist numerous qualitative studies on patient transitions, mathematical models addressing care transitions are limited. According to a survey on healthcare system models [69], although long waiting times in highly congested hospitals have received substantial attention, of the 88 literatures, only 30 explicitly modeled the interaction with upstream units. For instance, multivariate regression analysis is carried out in [70] to study delays on transfers from ICU to general care units. Simple queueing models are presented in [71] to study delays in ICU admissions and readmissions. However, most of such studies only focus on one specific transition without characterizing the interactions with other departments explicitly. In paper [72], separate queueing models are used for different units to study obstetric patient flow. Similarly, the patient flow within an obstetrics hospital is analyzed through a queueing model in [73] for bed balancing study without direct consideration of blocking.

However, analyzing a complex system without considering interdependencies can lead

to inaccurate and misleading results. Moreover, such errors can be amplified in highly congested hospital settings. Within a few studies encompassing multiple departments, queueing models are presented in reference [74] for patient flow between ED, ICU and medical units. In paper [75], each subnetwork is treated as an M/M/c queue and arrival and service rates are modified to consider blocking in allocating mental health resources. The patient's length of stay is estimated by decomposing the flow into simple tandem queues. Similarly, a tandem queue is used in [76] to study the blocking effect in a series of care in multiple departments. A closed queueing network to model patient flow in geriatric department is introduced in [77] by assuming that the system is always full and no feedback flow occurs. Paper [78] uses a finite capacity open queueing network to model patient flow in operative and post-operative units at a university hospital. Reference [46] analyzes the patient length of visit in an urgent care center using multi-class open queueing network model and incorporates fork/join queues. However, most prevailing studies either focus on single server stations or analyze multi-server stations under simple tandem networks (without feedback flow) or assuming that the first stage is never empty. In [79], Markov chain models are used to study transitions between ED, ICU and wards but is mostly applicable to small sized hospitals due to computational issues.

In summary, modeling and analysis of patient transitions encompassing multiple departments within hospitals are still largely unexplored. In addition, most existing literatures only focus on mean performance measures and hardly address variability issue. However, in healthcare delivery systems, addressing variability is necessary to solve many of the existing problems. As stated by the Institute of Medicine, hospitals need to understand the underlying variabilities in order to improve patient safety and quality while simultaneously reducing hospital waste and cost. To the best of our knowledge,

no study explicitly derived variability of the patient waiting times.

2.4 Optimal TJR Postoperative Care Management

As health systems are currently under strong economic pressure, there is an increasing emphasis on achieving cost-effectiveness in care throughout the medical community [80]. Regarding TJR, there is a sizable body of literature that investigates the cost-effectiveness of the joint replacement surgical procedure: according to a systematic review of cost-effectiveness analyses of THA and TKA [81], the cost-effectiveness of TJR compared to no surgery is estimated to be $10,402 per quality-adjusted life year (QALY) gained for THA and $12,000 to $20,000 per QALY for TKA, thus both procedures are considered highly cost-effective.

However, studies that address the postoperative intervention programs are rather scarce [82] where there exists a limited number of papers examining the efficacy of intervention before and after joint arthroplasty surgery [28]. Regarding the detailed intervention practices, as the intervention protocols vary widely in both the specific interventions used and the timeframes for their delivery, there exist two streams of literature of interest. The first stream is regarding the types of interventions, i.e., comparing intervention programs provided in different venues (e.g., physical therapy provided at home or in rehab facility). Several studies have compared intervention programs in hospital and at home after a total knee replacement and have found no group differences in the functional outcomes [83, 84]. Paper [85] shows a significant reduction in the cost of care by using home-based rehabilitation programs without compromising its quality. On the other hand, it is asserted in [82] that evidence of effects of different intervention programs following TKA is limited. Most of the existing studies examine selected PT

activities and investigate their effects on patient outcomes which are typically based on a treatment-control group design or controlled trials with small sample size [86].

The second stream of literature is related to investigating the intervention timeframes. Over the past decade, there has been a push towards implementing progressively earlier PT regimens (e.g., begin PT on postoperative day zero (POD 0)). Various studies have conducted systematic reviews and meta-analyses and concluded that intensive functional rehabilitation during the subacute recovery period after TKA is associated with improved functional capacity, pain intensity, and quality of life [87]. Paper [88] also shows that the initiation of early PT has improved short-term outcomes where PT intervention during the immediate postoperative period after THA is directly associated with less cost of care. However, the results from these studies are difficult to compare and interpret given variable methods for timing of PT initiation, duration, and frequency [89]. Moreover, with most literatures focusing on the immediate postoperative period, the effect of interventions during the more far end of the recovery progress is mainly unknown. As the timeframe of the CMS readmission reduction program for TJR is up to 90 days since discharge from the hospital, current literature lacks evidence in supporting decision making across the entire timeframe.

One of the analytical tools that have been utilized as an effective tool in such medical decision making problems is Markov decision process. MDP has been applied to various healthcare areas to develop tools that can guide healthcare professionals in making evidence-based decisions. Paper [90] investigates the problem of optimally timing a living-donor liver transplant by using an MDP model, and derives the structural properties of the optimal policy. Similarly, paper [91] formulates the optimal breast cancer biopsy decision model as an MDP to investigate the structural properties of the model to gain insights on how decisions are made, and paper [92] uses a partially observable

Markov decision processes (POMDP) to obtain an age and belief dependent optimal biopsy referral policy for prostate cancer. As such, the use of MDP models has been gaining increasing popularity in medical decision making problems.

Ideally, health professionals' decisions about appropriate postoperative care process should be aided by evidence-based clinical practice guidelines [82]. However, in TJR rehabilitation, the evidence is limited and inconsistent [87, 89]. Therefore, there is a need to establish optimal practice guidelines based on health professional expertise, data and evidence [27].

Chapter 3

Improving Access to Primary Care

3.1 Introduction

In this chapter, we introduce a Markov chain based modeling framework to assess the impact of implementing a new service model in primary care clinics. The systems under study (both current and the proposed new service model) are described in Section 3.2. Markov chain models for the current and future systems are introduced in Sections 3.3 and 3.4, respectively. Furthermore, to illustrate the applicability of the model, an application study of patient flow modeling at Dean East Clinic is introduced. Dean East Clinic, a primary and comprehensive care clinic, is part of Dean Health Systems which is one of the largest integrated healthcare delivery systems in the US. Finally, conclusions and managerial insights from the model are formulated in Section 3.5.

3.2 System Description

In many primary care clinics, the care services are divided into pods. Each pod consists of multiple exam rooms, and is handled by a provider group consisting of a number of physicians (MD) (or in some cases may include physician assistants (PA), nurse practitioners (NP), etc.) and medical assistants (MA). Typically, an MA is dedicated to support one provider. The patient flow in each exam room includes:

- MA escort and rooming: Once the patient is escorted to the exam room by the MA, the patient's weight, pulse, blood pressure, syndrome, medication and basic information are measured or recorded.

- Provider visit: The provider checks the patient's syndrome, diagnoses, and provides necessary prescription.

- MA visit: MA wraps up the visit, provides necessary medication such as wound care and/or immunization, and finally, discharges the patient.

In Dean East Clinic, each pod consists of four exam rooms and two providers (an MD and a PA). Each provider is in charge of two rooms and has a dedicated MA to support. We denote these two groups as MD team and PA team. In Figure 3.2, an illustration of patient flow in two exam rooms for a team of one provider and one MA is presented, where the circles refer to the services carried out by the provider or MA, and the rectangles represent the waiting states (i.e., buffer) in which patients are waiting for the next service due to unavailable resource.

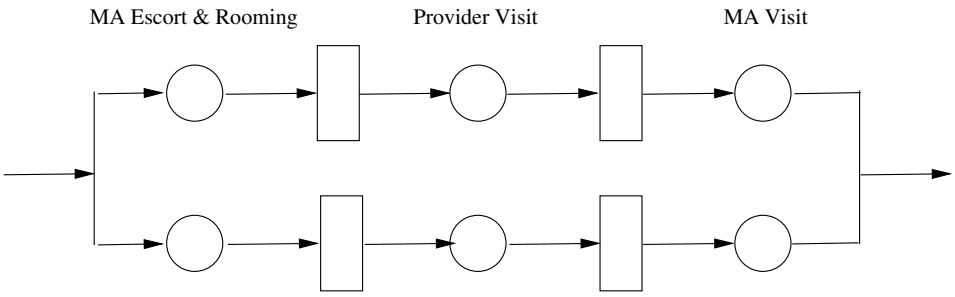

Figure 3.2: Current system patient flow

In addition to the direct face-to-face work with patients, once a patient's visit is over, providers need to input all relevant details of the visit into the EMR system. The ideal case is to input the information right after a patient's visit. However, due

to tight schedules, such inputs may not be finished promptly. Providers may need to work extra time on a daily basis and even on weekends to finish patient charting. These long gaps between the visits and data entry would affect the data quality and may lead to loss of necessary details. Even worse, such burdens may impact the provider-patient relationship and eventually the overall care quality. Thus, current work flow model lacks in ensuring timely access to care for the patients and better utilization of the providers' time. Moreover, relatively low utilization of MAs compared to that of providers implies that the workloads between providers and MAs are not balanced for the optimal efficiency and cost-effectiveness.

To improve system performance, a joint visit service model is proposed to redesign the work flow of both the provider and MA. In such practices, some portion of non-direct care workload is shifted from the provider to the MA. Instead of provider visiting the patient alone during the current provider visit phase, the MA will accompany the provider which is referred to as the joint visit phase. During this joint visit phase, while the provider focuses on the patient, the MA assists the provider by documenting necessary notes into the EMR system and thus conducting some portion of provider's non-direct care work. Upon finishing, the MA leaves for other patients or indirect care work, and the provider wraps up the visit. Afterwards, the MA provides necessary medication if needed, and finally discharges the patient. The patient flow of such a model is illustrated in Figure 3.3 for two exam rooms with one provider. Note that there is no waiting between the joint visit and the provider wrap-up as the provider conducts the two tasks consecutively.

Moreover, to further reduce provider's workload, MA wrap-up system has been proposed as well. As shown in Figure 3.4, once joint visit is done, provider leaves for other

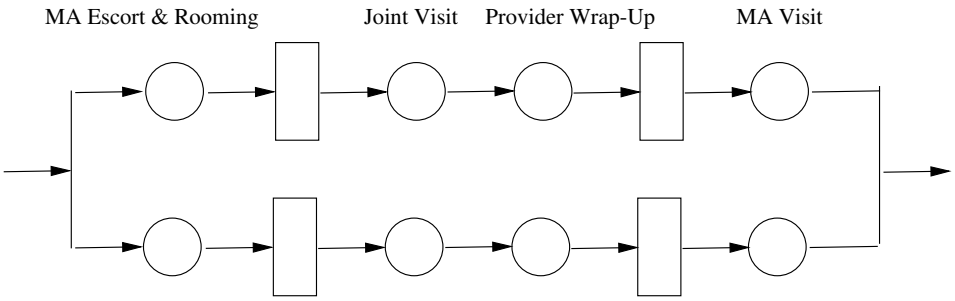

Figure 3.3: Joint visit model of one provider: Provider wrap-up

patients or indirect care work, and MA wraps up the service, performs additional medication and discharges the patient. As the MA conducts all these activities continuously, there is no waiting between joint visit and MA wrap-up, and MA wrap-up and MA visit can be grouped as a single activity.

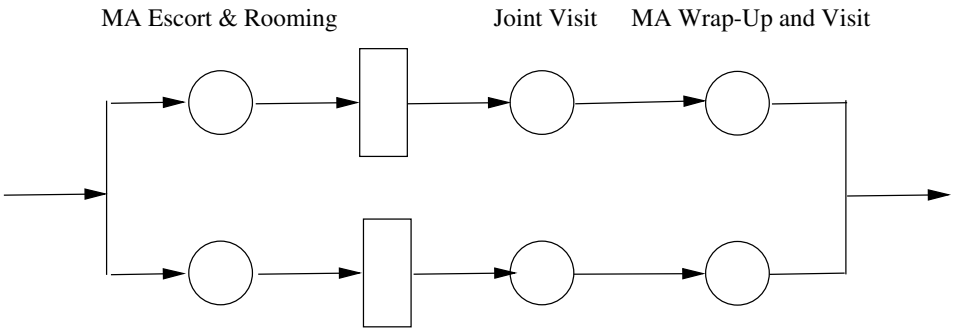

Figure 3.4: Joint visit model of one provider: MA wrap-up

The above systems represent one of the service models prevalent in many primary care clinics and the promising redesigns that are expected to improve the patient flow. Thus, studying these models could bring insights to the clinical managers who are planning to adopt such joint visit systems. In this chapter, we develop analytical methods to evaluate and compare the performance of these service models and provide recommendations for system redesign.

3.3 Current System

To study the patient flow, we first analyze the current dedicated visit system. Based on the current service setting at Dean East Clinic, a Markov Chain model is introduced and is further extended to a general distribution case. Then, the proposed joint visit systems with provider and MA wrap-ups are explored, and compared with the current system performance.

3.3.1 Dedicated Visit System

Markov Chain Model

As the two providers serve the patients independently, and the two MAs support the providers exclusively, we can analyze the provider teams separately. In other words, a model can be developed for each provider. We assume:

- Service times follow exponential distribution.
- There are always patients available for rooming, i.e., unlimited arrivals.
- Patient occupies an exam room during the entire visit. Thus, there can be only one patient per each exam room.
- All demands competing for the same resource (provider or MA) follow a non-preemptive priority rule, and MA visit has a higher priority than escort and rooming.

Now, using the MD team as an example, we can define the system states as $S = \{s_{\text{room}}, b_{\text{diag}}, s_{\text{diag}}, b_{\text{med}}, s_{\text{med}}\}$, where s_i represents the number of patients receiving service

by a resource (MD or MA), and b_j represents the number of patients who have finished service but are waiting for the next service due to unavailable resource at the next stage. In these waiting states, the current stage resource is released. Thus, s_i's and b_j's are represented by circles and rectangles in Figure 3.2, respectively. Note that 'room', 'diag', 'med' are used to denote rooming, provider visit (for diagnosis) and MA visit (for medication), respectively. Since a team consists of only one MA and one provider, the scenarios of two patients receiving the same service simultaneously or even though not the same service, being served by the same resource (e.g., MA rooming and MA visit are both served by MA) are infeasible. In addition, both patients cannot be in waiting states at the same time. Thus, the feasible states of the current system include:

$$S_1 = (1,0,0,0,0), \quad S_2 = (1,0,1,0,0), \quad S_3 = (1,0,0,1,0), \quad S_4 = (0,1,1,0,0),$$
$$S_5 = (0,0,1,0,1), \quad S_6 = (0,0,0,1,1), \quad S_7 = (0,0,0,0,1).$$

The state transitions and the corresponding balance equations can be derived (see Appendix A) and the system performance can be evaluated by solving them. Let P_i, $i = 1, \ldots, 7$, be the steady state probability of state S_i, and c_k be the service rate of care activity $k \in \{\text{room}, \text{diag}, \text{med}\}$. Denote TP^{exp} and LOV^{exp} as the throughput (number of patients leaving the system per hour) and average length of visit (total time spent in the exam room, i.e., in-room cycle time) of the Markov chain model, respectively, where subscript 'exp' indicates the Markovian assumption. In addition, let N be the average number of patients in the system, and ρ_j, j =MD, MA, be the utilization of the care giver (i.e., percentage of direct face-to-face encounter time with patients). We obtain:

Proposition 3.1 *In the current systems with exponential service times, the system performance can be calculated as:*

$$\begin{aligned} TP^{\exp} &= c_{\text{med}}(P_5 + P_6 + P_7), \\ N &= P_1 + P_7 + 2\sum_{i=2}^{6} P_i, \\ LOV^{\exp} &= \frac{N}{TP} = \frac{P_1 + P_7 + 2\sum_{i=2}^{6} P_i}{c_{\text{med}} \sum_{i=5}^{7} P_i}, \\ \rho_j &= \begin{cases} P_2 + P_4 + P_5, & j = \text{MD}, \\ \sum_{i=1, i\neq 4}^{7} P_i, & j = \text{MA}. \end{cases} \end{aligned} \tag{3.1}$$

Non-Marovian Case

The above mentioned model assumes exponential service times. However, as this assumption may not be valid in practice, non-exponential service times are investigated to relax this assumption. To do so, we first verify that the system performance is practically independent of the distribution type, but mainly depends on the mean and CV of the service times. If that is the case, then an empirical formula can be proposed to approximate the performance under non-exponential distributions.

First, we investigate whether the distribution type affects system performance significantly or not. In other words, is the model practically independent of distribution type? To identify the dependency, extensive simulation experiments are conducted using various distribution types including Gamma (G), Lognormal (L), Weibull (W), and Mix (M), which refers to a mixed distribution where service time is randomly chosen from the three distributions (G, L, and W). In simulation experiments, the model is setup to follow the patient flow depicted in Figure 3.2, and the same assumptions introduced in Markov chain model (except exponential service time distributions). The parameters

are selected based on the data collected at the clinic. 1000 time units of warm up period is assumed and the remaining 100,000 time units are used as data collection period. Such a setting is used for all simulation studies throughout the chapter. To quantify the differences in LOV among different distributions, we introduce

$$\delta = \frac{\max_j LOV_i - \min_j LOV_j}{\frac{1}{4}\sum_j LOV_j}, \quad j \in \{\mathrm{W}, \mathrm{G}, \mathrm{L}, \mathrm{M}\},$$

where LOV_j is the LOV using distribution j. In addition, to investigate the impact of variability on the differences among various distributions, we vary the coefficient of variation(CV) from 0.2 to 1 in increments of 0.2. Note that only CVs smaller or equal to 1 are considered, since in most healthcare systems, CV tends to be less than 1, implying that the longer the service has been carried out, the higher the probability the service will be finished [93]. Among all the experiments we carried out, the δ's for both the MD and PA under all CVs are usually within 1%, and the maximum is less than 2%. Thus, this result indicates that the LOV is practically independent of distribution type. Indeed, such independence is also observed in other healthcare system studies (e.g., [50, 53, 94, 51, 49]) and manufacturing systems as well [93].

Due to this independence, an empirical formula can be pursued to analyze the general distribution case. By varying the CV from 0.2 to 1, it is identified that LOV is monotonically increasing with respect to CV where a quadratic relationship is observed between CV and LOV. Thus, the following empirical formula is proposed:

$$LOV^{\mathrm{non-exp}} = (LOV^{\mathrm{exp}} - LOV^0) \cdot (CV^{\mathrm{eff}})^2 + LOV^0, \tag{3.2}$$

where $LOV^{\mathrm{non-exp}}$ represents the length of visit under CV between 0 and 1, and LOV^{exp} refers to the result obtained from the Markov chain model (i.e., exponential case when $CV = 1$). In addition, LOV^0 implies there is no variability ($CV = 0$) in the service time, so that the length of visit can be determined by either the total time for a resource

type (MD or MA) to serve all exam rooms or the total time it takes a patient to go through all the services, whichever is larger, i.e.,

$$LOV^0 = \max\Big\{\frac{2}{c_{\text{room}}} + \frac{2}{c_{\text{med}}}, \frac{2}{c_{\text{diag}}}, \sum_{k\in\{\text{room,diag,med}\}} \frac{1}{c_k}\Big\}. \tag{3.3}$$

Moreover, $(CV^{\text{eff}})^2$ is defined as the overall weighted CV squares,

$$(CV^{\text{eff}})^2 = \frac{\sum_{k\in\{\text{room,diag,med}\}} \frac{1}{c_k}\cdot(cv_k)^2}{\sum_{k\in\{\text{room,diag,med}\}} \frac{1}{c_k}}, \tag{3.4}$$

where cv_k represents the CV of service s_k, $k \in \{\text{room}, \text{diag}, \text{med}\}$.

Accuracy

To investigate the accuracy of the proposed models, validations by both simulations and the data collected in Dean East Clinic are carried out. Through multiple observations, interviews and analysis of EMR data, the average service times (i.e., inverse of service rates) are estimated as shown in Table 3.2.

Table 3.2: Service times of current system

	MA escort & rooming	Provider visit	MA visit
$1/c_k$(min)	8.78	31.204	2.892
cv_k	0.389	0.401	0.45

First, using Proposition 3.1 and empirical formulas (3.2)-(3.4), we obtain the patient average length of visit as 63.2 minutes while the collected data is 58 minutes, thus resulting in a difference of 8.97%. As no-show of patients or open appointment slots are not considered in the analytical model, and considering the discrepancy in the observed data, such a model provides an acceptable accuracy.

Secondly, we investigate the impact of variability on LOV by changing the CVs and comparing the results obtained from our analytical models with simulations. For each service, CV is randomly chosen between [0,1]. Define

$$\epsilon^{\text{non-exp}} = \frac{|LOV^{\text{non-exp}} - LOV^{\text{sim}}|}{LOV^{\text{sim}}} \cdot 100\%, \tag{3.5}$$

where $LOV^{\text{non-exp}}$ and LOV^{sim} denote the results from the analytical model and simulations, respectively. Here, simulation service time distributions are randomly selected among Gamma, Lognormal, Weibull, and Mix where the parameters for each distribution is calculated using the average service time shown in Table 3.2 and random CV. As shown in Table 3.3, the differences between LOVs from the empirical formula and simulations are small (well below 1.67%), which implies that empirical formulas (3.2)-(3.4) provide close estimates.

Thirdly, to see whether the aforementioned models work under different settings, dozens of experiments are conducted with randomly selected service rates and random CVs in the range of [0,1]. Using the minimum and maximum values of the collected data as the lower and upper bounds of sets following uniform distributions, the service rates are defined as:

$$\begin{aligned} \text{MA rooming:} \quad & \frac{1}{c_{\text{room}}} \in U(4, 14), \\ \text{Provider visit:} \quad & \frac{1}{c_{\text{diag}}} \in U(8, 47), \\ \text{MA visit:} \quad & \frac{1}{c_{\text{med}}} \in U(3.2, 9.4). \end{aligned}$$

Simulation experiments are conducted in a similar fashion as before, except now the average service times are randomly chosen from the uniform distributions. Results of 20 random experiments are shown in Table 3.4, where the relative differences (defined similarly as (3.5)) in throughput, LOV, and provider utilizations are listed. Again, the

Table 3.3: Accuracy of current system model: general case

$(CV^{\text{eff}})^2$	$LOV^{\text{non-exp}}$	LOV^{sim}	$\epsilon^{\text{non-exp}}$
0.074	62.794	62.511	0.453%
0.124	63.054	62.612	0.707%
0.259	63.757	63.567	0.299%
0.344	64.200	64.078	0.189%
0.466	64.834	63.853	1.536%
0.607	65.567	64.652	1.415%
0.627	65.673	65.040	0.974%
0.645	65.766	65.448	0.485%
0.722	66.165	65.079	1.669%
0.938	67.291	66.299	1.496%

average differences between simulation and analytical model under random parameters are minimal. Therefore, we claim that Proposition 3.1 and empirical formulas (3.2)-(3.4) provide a close estimate of the system performance.

Table 3.4: Accuracy of the current system model with randomized data

	Throughput	Length of visit	Utilization
$\epsilon^{\text{non-exp}}$	0.555%	0.554%	0.063%

3.3.2 Joint Visit System with Two MAs

Using a similar approach, we study the proposed joint visit system with two MAs (in both scenarios of provider wrap-up and MA wrap-up). Same assumptions are used as the current dedicated system except one additional assumption regarding the joint visit priority. As the purpose of implementing a joint visit system is to make more efficient use of the providers, we need to minimize the time provider spends waiting for an available MA in order to start the joint visit service. This in turn implies that the highest priority should be assigned to joint visit among all the care activities. Thus for MA, priority is the highest for joint visit and lowest for escort and rooming. The detailed state space definition, balance equations, and calculation formulas are presented in Appendix A. The results are shown in Table 3.5 where Throughput is presented in (1/hr) and LOV in (min).

As one can see, the utilization of providers has been reduced while the utilization of MAs has been increased substantially. Thus, by shifting workloads from providers to MAs, providers' workload burden has been relieved. However, as patient needs to be seen by both the provider and MA, either resource being too highly utilized results in an

unbalanced system, hence decreasing patient throughput. Implementation of joint visit only shifts the bottleneck of the system from the provider to the MA without improving the patient flow. Therefore, to fully benefit from implementing the joint visit system, balanced workload among the resources needs to be achieved. One way to achieve this is to recruit an additional MA, and analysis of such a scenario is presented in Section 3.4.

Table 3.5: Comparison between current system and joint visit system with two MAs

	Throughput	LOV	Provider utilization	MA utilization
Current	1.759	63.200	91.454%	34.208%
Joint/Provider wrap-up	1.616	73.200	81.541%	98.247%
Joint/MA wrap-up	1.501	39.972	66.972%	100%

Remark 3.1 The reduction of LOV in joint visit with MA-wrap up is due to the way how LOV is defined. LOV, which is an in-room cycle time, only includes the waiting times inside exam rooms, thus it does not include the waiting time for rooming. In joint visit with MA-wrap up system, MA is required for all the services carried out in a patient's visit. Thus, for current system where there is only one MA per team, as rooming a new patient has the lowest priority among MA's tasks, MA would not work on a new patient until current patient finishes all services and leaves clinic. In this sense, once a patient is roomed, there is no waiting time and LOV would simply be the sum of the service times. Note that this does not mean a patient experiences zero waiting time. Patients may need to wait to get roomed, but this waiting time is not included in the current LOV definition. □

3.4 Future Joint Visit System

Since directly implementing joint visit system leads to a lower throughput and longer waiting time, a future solution of adding one additional MA in each pod is proposed. Therefore, for each pod, there will be three MAs supporting two providers. In addition, instead of dedicated MA system, a float MA system is proposed, that is, all three MAs will work as a team to support both providers based on a first come (service request) first serve policy. Below, both provider wrap-up and MA wrap-up cases are studied under this new setting.

3.4.1 Joint Visit System with Three MAs and Provider Wrap-up

System States

Now with the float MA system, the performance of the providers cannot be analyzed separately as before. A straightforward approach is to develop a model for the entire pod (2 provider, 3 MA, 4 exam room system). Although one can do this by explicitly listing all the scenarios and defining the system states, there will be a dimensionality issue if the system size is increased, such as increasing the number of exam rooms or adding additional MAs or providers to the pod. One other limitation of developing a model for the entire pod is that it becomes more difficult to keep track of the system when providers' performance substantially differ. Specifically for Dean clinic, the patient pool itself is different between MD and PA. While PA patients usually have regular office visits, MD patients require more comprehensive examination. Also, even when the patient pool is similar, large performance variance is typical in healthcare settings, so generalizing

provider visit as a single service distribution may introduce more discrepancy with the real world. In this sense, to keep track of which provider is working on which patient, additional system states may need to be introduced. Therefore, an alternative approach is pursued to avoid such overcomplexity issues.

In each pod, the two providers serve the patients independently with different service rates. Typically, more complex patients are seen by the MD, thus resulting in lower service rates. Following the sequence of stages as depicted in Figure 3.3, system states for one provider can be defined as follows (where 'joint' and 'wrap' indicate joint visit and provider wrap-up, respectively):

$$S = \{s_{\text{room}}, b_{\text{joint}}, s_{\text{joint}}, s_{\text{wrap}}, b_{\text{med}}, s_{\text{med}}\}. \tag{3.6}$$

As MD and PA have different service rates and have no overlapping patients, an intuitive approach is to analyze the two teams separately. However, since three MAs are shared among the providers, MD team and PA team are no longer independent. Instead, one team's availability of MA depends on the other team's state. Specifically, there exist two scenarios of allocating MAs to one provider (e.g., MD):

(1) Two MAs are serving both patients of PA, thus only one MA is available to serve MD patients.

(2) Only one MA is serving PA patients, thus two MAs are available to serve both patients of MD.

Thus, the MA's availability for the MD is dependent on that for the PA, which is unknown. To accommodate this unknown availability and to reduce the dimension of the Markov chain, an iterative method is introduced. The idea is to first assume there are two MAs dedicated to one provider (i.e., 4 MAs in the pod). Now we can analyze each system separately with one provider, two MAs, and two exam rooms. In this case,

the system states are defined as

$$S_1 = (2,0,0,0,0,0), \quad S_2 = (1,0,1,0,0,0), \quad S_3 = (0,1,1,0,0,0),$$
$$S_4 = (0,1,0,1,0,0), \quad S_5 = (0,0,1,0,0,1), \quad S_6 = (1,0,0,1,0,0),$$
$$S_7 = (1,0,0,0,0,1), \quad S_8 = (0,0,0,0,0,2), \quad S_9 = (0,0,0,1,0,1).$$

With the system states described as above, the state transitions can be derived as well as the balance equations (see Figure A.4 and equations (A.5) in Appendix A). Again, let P_j, $j = 1,\ldots,9$, be the steady state probability of state S_j, and c_k, $k \in \{\text{room}, \text{joint}, \text{wrap}, \text{med}\}$, be the service rate of care activity k. By solving these equations we can obtain the steady state probabilities P_j, $j = 1,\ldots,9$. We denote this calculation as operator $\Phi_j(\cdot)$.

$$P_j = \Phi_j(c_{\text{room}}, \ldots, c_{\text{med}}), \quad j = 1,\ldots,9.$$

Now using these derived P_j's, performance metrics can be evaluated. The throughput of the system can be viewed as the discharge rate from the last service.

$$TP^{\text{exp}} = c_{\text{med}}(P_5 + P_7 + 2P_8 + P_9). \tag{3.7}$$

However since such an assumption (two MAs dedicated to a provider) will lead to overestimating the throughput of the system, we adjust the service rates of MAs to take into account the probability of MA not being available under certain circumstances (scenario (1)).

Iteration Procedure

To adjust the MA's service rate, define α^i, i =MD, PA, as the probability that both MAs are busy for provider i, which coincide with states S_1, S_2, S_5, S_7, and S_8. Then, we

update the MA's service rate by accounting for the probability α of the other provider. Also, since there exist more than one MA in the system, we introduce a discount factor β. Here, discount factor is introduced to prevent over reducing MA's service rates. As we assume two MA's per team, $\beta = 1/\text{number of MAs} = 0.5$ is selected. Thus, using $\alpha \cdot \beta$, we decrease the service rate of MA visit c_{med} to $\widetilde{c}_{\text{med}}$. The rationale of such a decrease is that when all MAs are being fully utilized, escorting and rooming can only be done when the current patient whom MA is visiting leaves the clinic. Thus, decreasing c_{med} to $\widetilde{c}_{\text{med}}$ corresponds to MA being unavailable to escort and room the next patient when all MAs of the other provider are busy.

$$\widetilde{c}_{\text{med}}^{\text{MD}} = c_{\text{med}}^{\text{MD}}(1 - \alpha^{\text{PA}} \cdot \beta),$$

$$\widetilde{c}_{\text{med}}^{\text{PA}} = c_{\text{med}}^{\text{PA}}(1 - \alpha^{\text{MD}} \cdot \beta).$$

Using the new $\widetilde{c}_{\text{med}}^{i}$, i=MD, PA, the steady state probabilities are recalculated as $\widetilde{P}_j^i$, $j = 1, \ldots, 9$.

$$\widetilde{P}_j^{\text{MD}} = \Phi_j(c_{\text{room}}, c_{\text{joint}}, c_{\text{wrap}}, \widetilde{c}_{\text{med}}^{\text{MD}}),$$

$$\widetilde{P}_j^{\text{PA}} = \Phi_j(c_{\text{room}}, c_{\text{joint}}, c_{\text{wrap}}, \widetilde{c}_{\text{med}}^{\text{PA}}).$$

Then, the performance measures can be updated.

$$TP^{\text{MD}} = \widetilde{c}_{\text{med}}^{\text{MD}}(\widetilde{P}_5^{\text{MD}} + \widetilde{P}_7^{\text{MD}} + 2\widetilde{P}_8^{\text{MD}} + \widetilde{P}_9^{\text{MD}}),$$

$$TP^{\text{PA}} = \widetilde{c}_{\text{med}}^{\text{PA}}(\widetilde{P}_5^{\text{PA}} + \widetilde{P}_7^{\text{PA}} + 2\widetilde{P}_8^{\text{PA}} + \widetilde{P}_9^{\text{PA}}).$$

However, as probabilities α^{MD} and α^{PA} are unknown, a recursive procedure is introduced to update the service rates during each iteration. Formally, such a procedure can be described as follows:

Procedure 3.1

Step 1: Initialization.

- *Step 1.1: Calculate the initial steady state probabilities P_i, $i = 1, \ldots, 9$, using operator $\Phi(\cdot)$.*

$$\widetilde{P}_j^{i,0} = \Phi_j(c_{\text{room}}^i, c_{\text{joint}}^i, c_{\text{warp}}^i, \widetilde{c}_{\text{med}}^{i,0}), \quad j = 1, \ldots, 9, \quad i = \text{MD}, \text{PA},$$

where superscript '0' denotes iteration 0, and

$$\widetilde{c}_{\text{med}}^{i,0} = c_{\text{med}}^i, \quad i = \text{MD}, \text{PA}.$$

- *Step 1.2: Calculate the throughput of each provider.*

$$TP^{i,0} = \widetilde{c}_{\text{med}}^{i,0}(\widetilde{P}_5^{i,0} + \widetilde{P}_7^{i,0} + 2\widetilde{P}_8^{i,0} + \widetilde{P}_9^{i,0}), \quad i = \text{MD}, \text{PA}.$$

- *Step 1.3: Calculate probability α.*

$$\alpha^{i,0} = \widetilde{P}_1^{i,0} + \widetilde{P}_2^{i,0} + \widetilde{P}_5^{i,0} + \widetilde{P}_7^{i,0} + \widetilde{P}_8^{i,0}, \quad i = \text{MD}, \text{PA}.$$

In addition, let $n = 1$, where n is iteration number.

Step 2: Updating.

- *Step 2.1: Update service rate c_{med}, steady state probability P_i^{MD}, and probability α^{MD} for MD.*

$$\begin{aligned}
\widetilde{c}_{\text{med}}^{\text{MD},n} &= c_{\text{med}}^{\text{MD},0}(1 - \alpha^{\text{PA},n-1} \cdot \beta), \\
\widetilde{P}_j^{\text{MD},n} &= \Phi_j(c_{\text{room}}^{\text{MD}}, c_{\text{joint}}^{\text{MD}}, c_{\text{wrap}}^{MD}, \widetilde{c}_{\text{med}}^{MD,n}), \quad j = 1, \ldots, 9, \\
\alpha^{\text{MD},n} &= \widetilde{P}_1^{\text{MD},n} + \widetilde{P}_2^{\text{MD},n} + \widetilde{P}_5^{\text{MD},n} + \widetilde{P}_7^{\text{MD},n} + \widetilde{P}_8^{\text{MD},n}.
\end{aligned} \tag{3.8}$$

- *Step 2.2: Update service rate* c_{med}*, steady state probability* P_i^{PA}*, and probability* α^{PA} *for PA.*

$$\begin{aligned}
\widetilde{c}_{\text{med}}^{\text{PA},n} &= c_{\text{med}}^{\text{PA},0}(1-\alpha^{\text{MD},n}\cdot\beta),\\
\widetilde{P}_j^{\text{PA},n} &= \Phi_j(c_{\text{room}}^{\text{PA}}, c_{\text{joint}}^{\text{PA}}, c_{\text{wrap}}^{\text{PA}}, \widetilde{c}_{\text{med}}^{\text{PA},n}), \quad j=1,\ldots,9,\\
\alpha^{\text{PA},n} &= \widetilde{P}_1^{\text{PA},n}+\widetilde{P}_2^{\text{PA},n}+\widetilde{P}_5^{\text{PA},n}+\widetilde{P}_7^{\text{PA},n}+\widetilde{P}_8^{\text{PA},n}.
\end{aligned} \tag{3.9}$$

- *Step 2.3: Update throughput for each provider.*

$$TP^{i,n} = \widetilde{c}_{\text{med}}^{i,n}\Big(\widetilde{P}_5^{i,n}+\widetilde{P}_7^{i,n}+2\widetilde{P}_8^{i,n}+\widetilde{P}_9^{i,n}\Big), \quad i=\text{MD},\text{PA}. \tag{3.10}$$

Step 3: Comparison.

Let $\delta = 10^{-4}$ *be the stopping criteria. If*

$$\max_i |TP^{i,n}-TP^{i,n-1}| < \delta, \quad i=\text{MD},\text{PA},$$

then stop the iteration, and let

$$TP^i = TP^{i,n}, \quad i=\text{MD},\text{PA}.$$

Otherwise, let $n=n+1$*, and return to Step 2.*

Proposition 3.2 *Given the joint visit system with three MAs and provider wrap-up, Procedure 3.1 is convergent, i.e., the following limits exist:*

$$\begin{aligned}
TP^i &= \lim_{n\to\infty} TP^{i,n}, \quad i=\text{MD},\text{PA},\\
\widetilde{P}_j^i &= \lim_{n\to\infty} \widetilde{P}_j^{i,n}, \quad i=\text{MD},\text{PA}, \quad j=1,\ldots,9,\\
\widetilde{c}_{\text{med}}^i &= \lim_{n\to\infty} \widetilde{c}_{\text{med}}^{i,n}, \quad i=\text{MD},\text{PA}.
\end{aligned} \tag{3.11}$$

Proof: See Appendix A. ■

In addition to throughput, other performance measures such as length of visit and average number of patients in the system are of interest as well. However, since we assume two dedicated MAs per provider, the number of patients in the system will always be two, which is certainly not true. In real case, a new patient may not be roomed right after a patient leaves the clinic due to the need for MA to serve a higher priority patient in other exam rooms. Thus we not only need to adjust the service rate of MAs, but also the average number of patients in the system.

Proposition 3.3 *The average number of patients in the joint visit system can be evaluated as*

$$N^i = 1 + \sqrt{1 - TP^i \cdot w^i}, \quad i = \text{MD}, \text{PA}, \tag{3.12}$$

where TP^i is obtained from (3.11) and

$$w^i = \frac{1}{\widetilde{c}^i_{\text{med}}} - \frac{1}{c^i_{\text{med}}}, \quad i = \text{MD}, \text{PA}. \tag{3.13}$$

Proof: See Appendix A. ■

Using TP^i and N^i, i =MD, PA, the patient LOV under Markovian assumption can be obtained.

$$LOV^{i,\text{exp}} = \frac{N^i}{TP^i}, \quad i = \text{MD}, \text{PA}. \tag{3.14}$$

In addition, the utilizations of the providers and MAs can be evaluated.

$$\rho_i = \widetilde{P}^i_2 + \widetilde{P}^i_3 + \widetilde{P}^i_4 + \widetilde{P}^i_5 + \widetilde{P}^i_6 + \widetilde{P}^i_9, \quad i = \text{MD}, \text{PA},$$
$$\rho_{\text{MA}} = \frac{\sum_{i=\text{MD,PA}}[2(\widetilde{P}^i_1 + \widetilde{P}^i_2 + \widetilde{P}^i_5 + \widetilde{P}^i_7 + \widetilde{P}^i_8) + \widetilde{P}^i_3 + \widetilde{P}^i_6 + \widetilde{P}^i_9]}{3},$$

where the MA utilization considers the number of working (non-idle) MAs in the defined state.

Proposition 3.2 provides a method to estimate the length of visit for the joint visit system with three MAs and provider wrap-up under the exponential assumption. In case of general service times, empirical formula (3.2) is used where $\forall i = MD, PA$,

$$
\begin{aligned}
LOV^{i,\text{non-exp}} &= (LOV^{i,\text{exp}} - LOV^{i,0}) \cdot CV^{i,\text{eff}} + LOV^{i,0}, \\
LOV^{i,0} &= \max\left\{ \frac{2}{3}\left(\sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^{\text{MD}}} + \sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^{\text{PA}}} \right) \right. \\
&\quad \left. \cdot \min_{l=\text{MA,PA},l\neq i} \left(1, \frac{\sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^i}}{\sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^l}} \right), \frac{2}{c^i_{\text{joint}}} + \frac{2}{c^i_{\text{wrap}}}, \sum_{k\in\{\text{room,joint,wrap,med}\}} \frac{1}{c_k^i} \right\}, \\
CV^{i,\text{eff}} &= \frac{\sum_{k\in\{\text{room,joint,wrap,med}\}} \frac{1}{c_k^i} \cdot (cv_k^i)^2}{\sum_{k\in\{\text{room,joint,wrap,med}\}} \frac{1}{c_k^i}}.
\end{aligned}
\tag{3.15}
$$

Accuracy Analysis

To investigate the accuracy of the proposed iterative procedure, we validate by using the data collected at Dean East Clinic. Since the joint visit system has not been implemented, estimates are obtained by interviewing the providers and MAs as shown in Table 3.6.

Table 3.6: Parameters of joint visit system

	Rooming		Joint Visit		Wrap-up	MA Visit
Provider team	MD	PA	MD	PA	MD & PA	MD & PA
c_k (1/min)	0.147	0.209	0.037	0.051	0.286	0.346
cv_k	0.389	0.741	0.401	0.397	0.45	0.45

Using these service rates, we compare the results of our iterative method and simulation. Let

$$\epsilon^{i,\exp} = \frac{|LOV^{i,\exp} - LOV^{i,\text{sim}}|}{LOV^{i,\text{sim}}} \cdot 100\%, \quad i = \text{MD}, \text{PA}.$$

As illustrated in Table 3.7(a), all errors are within 1.65%.

Table 3.7: Accuracy of joint visit system with 3 MAs and provider wrap-up: Markov chain case

	Throughput(1/hr)		Length of Visit(min)	
Provider team	MD	PA	MD	PA
Simulation	1.888	2.475	62.758	47.430
Model	1.920	2.497	62.284	47.824
$\epsilon^{\exp}$	1.648%	0.894%	0.755%	0.830%

(a) Collected data setting

	Throughput		Length of Visit	
Provider team	MD	PA	MD	PA
$\epsilon^{\exp}$	2.094%	0.932%	0.769%	0.500%

(b) Randomized data setting

To further validate the developed model under different settings, randomly generated service rates are used in dozens of experiments. Again, uniform distributions are used to quantify the service times where the ranges of the service times are presented below. Note that here Markovian assumption still holds. Based on more than two dozen random experiments, as shown in Table 3.7(b), the average difference between simulation and

analytical model in the Markovian scenario is less than 2.1%.

$$\text{Joint visit:} \quad \frac{1}{c^i_{\text{joint}}} \in \begin{cases} U(8,40), & i = \text{MD}, \\ U(8,33), & i = \text{PA}, \end{cases}$$

$$\text{Wrap-up:} \quad \frac{1}{c^i_{\text{wrap}}} \in U(2,5), \quad i = MD, PA,$$

$$\text{MA visit:} \quad \frac{1}{c_{\text{med}}} \in U(1.2, 4.4).$$

Then, to check the robustness under varying CVs, estimated service rates are used but with random CVs in the range of [0,1]. General distributions are assumed for service times in simulations, which are randomly selected from {L,G,W,M}. Comparison results for MD patients and PA patients are shown in Table 3.8, where the errors are less than 1.61%.

Furthermore, for general scenarios, both service rates and CVs are randomly selected. By randomly selecting the service time distribution from {L,G,W,M}, and its CV from 0 to 1, we compare the results of empirical formulas with simulations. The results indicate that the differences in LOVs of MD and PA patients are 0.963% and 0.954%, respectively which validates that the proposed empirical formula can be used to evaluate system performance.

3.4.2 Joint Visit System with Three MAs and MA Wrap-up

Similar to the provider wrap-up case, when there are one provider, two MAs, and two exam rooms, the system states can be defined as $S = (s_{\text{room}}, b_{\text{joint}}, s_{\text{joint}}, s_{\text{med}})$, where s_{med} denotes the MA wrap-up & visit activity. A total of 6 states can be obtained:

$$S_1 = (2,0,0,0), \quad S_2 = (1,0,1,0), \quad S_3 = (0,1,1,0),$$

$$S_4 = (1,0,0,1), \quad S_5 = (0,0,1,1), \quad S_6 = (0,0,0,2).$$

Table 3.8: Accuracy of joint visit system with 3 MAs and provider wrap-up: general case

$(CV^{\text{eff}})^2$	$LOV^{\text{non-exp}}$	LOV^{sim}	$\epsilon^{\text{non-exp}}$
0.002	60.543	61.305	1.243%
0.161	60.882	60.952	0.114%
0.226	61.023	60.767	0.420%
0.315	61.212	61.392	0.292%
0.385	61.360	61.583	0.362%
0.395	61.382	61.836	0.734%
0.451	61.502	61.774	0.440%
0.496	61.597	61.595	0.003%
0.733	62.101	62.076	0.040%
0.934	62.531	62.379	0.244%

(a) MD patients

$(CV^{\text{eff}})^2$	$LOV^{\text{non-exp}}$	LOV^{sim}	$\epsilon^{\text{non-exp}}$ (%)
0.029	46.538	45.980	1.215%
0.126	46.667	46.015	1.416%
0.216	46.786	46.046	1.606%
0.222	46.794	46.370	0.915%
0.286	46.879	46.344	1.154%
0.308	46.908	46.222	1.484%
0.360	46.977	46.648	0.705%
0.438	47.080	46.694	0.828%
0.532	47.204	46.720	1.035%
0.970	47.784	47.225	1.183%

(b) PA patients

Again the state transitions and balance equations can be derived and solving them results in the evaluation (clearly, an overestimate) of system throughput. Compared with the case of provider wrap-up, formulas for calculating throughput and α are changed.

$$\alpha^i = P_1^i + P_2^i + P_4^i + P_5^i + P_6^i, \quad i = \text{MD}, \text{PA},$$
$$TP^i = c_{\text{med}}^i (P_4^i + P_5^i + 2P_6^i), \quad i = \text{MD}, \text{PA}. \tag{3.16}$$

Define operator $\Psi_i(\cdot)$ as the calculation process needed to obtain the steady state probabilities. With these modifications, Procedure 3.2 is introduced to estimate the system performance.

Procedure 3.2

Step 1: Initialization. For i =MD, PA,

$$\widetilde{P}_j^{i,(0)} = \Psi_j(c_{\text{room}}^i, c_{\text{joint}}^i, \widetilde{c}_{\text{med}}^{i,0}), \quad j = 1, \ldots, 6,$$
$$TP^{i,0} = \widetilde{c}_{\text{med}}^{i,0} (\widetilde{P}_4^{i,0} + \widetilde{P}_5^{i,0} + 2\widetilde{P}_6^{i,0}),$$
$$\alpha^{i,0} = \widetilde{P}_1^{i,0} + \widetilde{P}_2^{i,0} + \widetilde{P}_4^{i,0} + \widetilde{P}_5^{i,0} + \widetilde{P}_6^{i,0},$$

again superscript '0' denotes iteration 0, and

$$\widetilde{c}_{\text{med}}^{i,0} = c_{\text{med}}^i, \quad i = \text{MD}, \text{PA}.$$

In addition, let $n = 1$.

Step 2: Updating.

$$
\begin{aligned}
\widetilde{c}_{\text{med}}^{\text{MD},n} &= c_{\text{med}}^{\text{MD},0}(1-\alpha^{\text{PA},n-1}\cdot\beta),\\
\widetilde{P}_j^{\text{MD},n} &= \Psi_j(c_{\text{room}}^{\text{MD}}, c_{\text{joint}}^{\text{MD}}, \widetilde{c}_{\text{med}}^{\text{MD},n}), \quad j=1,\ldots,6,\\
\alpha^{\text{MD},n} &= \widetilde{P}_1^{\text{MD},n}+\widetilde{P}_2^{\text{MD},n}+\widetilde{P}_4^{\text{MD},n}+\widetilde{P}_5^{\text{MD},n}+\widetilde{P}_6^{\text{MD},n},\\
\widetilde{c}_{\text{med}}^{\text{PA},n} &= c_{\text{med}}^{\text{PA},0}(1-\alpha^{\text{MD},n}\cdot\beta),\\
\widetilde{P}_j^{\text{PA},n} &= \Psi_j(c_{\text{room}}^{\text{PA}}, c_{\text{joint}}^{\text{PA}}, \widetilde{c}_{\text{med}}^{\text{PA},n}), \quad j=1,\ldots,6,\\
\alpha^{\text{PA},n} &= \widetilde{P}_1^{\text{PA},n}+\widetilde{P}_2^{\text{PA},n}+\widetilde{P}_4^{\text{PA},n}+\widetilde{P}_5^{\text{PA},n}+\widetilde{P}_6^{\text{PA},n},\\
TP^{i,n} &= \widetilde{c}_{\text{med}}^{i,n}\Big(\widetilde{P}_4^{i,n}+\widetilde{P}_5^{i,n}+2\widetilde{P}_6^{i,n}\Big), \quad i=\text{MD},\text{PA}.
\end{aligned}
\tag{3.17}
$$

where $\Psi_j(\cdot)$ *denotes the calculation to obtain the steady state probabilities* P_j, $j=1,\ldots,6$.

Step 3: Comparison. If

$$\max_i |TP^{i,n}-TP^{i,n-1}|<\delta, \quad i=\text{MD},\text{PA},$$

then stop the iteration, and let

$$TP^i = TP^{i,n}, \quad i=\text{MD},\text{PA}.$$

Otherwise, let $n=n+1$, *and return to Step 2.*

Analogously, we can show that such a procedure is also convergent.

Proposition 3.4 *Given the joint visit system with three MAs and MA wrap-up, Procedure 3.2 is convergent. Thus, for i =MD, PA,*

$$
\begin{aligned}
TP^i &= \lim_{n\to\infty} TP^{i,n},\\
\widetilde{P}_j^i &= \lim_{n\to\infty} \widetilde{P}_j^{i,n}, \quad j=1,\ldots,6,\\
\widetilde{c}_{\text{med}}^i &= \lim_{n\to\infty} \widetilde{c}_{\text{med}}^{i,n}.
\end{aligned}
\tag{3.18}
$$

In addition,

$$N^i = 1 + \sqrt{1 - TP^i \cdot \left(\frac{1}{\widetilde{c}^i_{\text{med}}} - \frac{1}{c^i_{\text{med}}}\right)},$$
$$LOV^{i,\exp} = \frac{N^i}{TP^i}. \tag{3.19}$$

Proof: The proof is similar to that of Proposition 3.2. ∎

Using Proposition 3.4, empirical formula (3.2) can be used to estimate LOV under general service times. Note that in this case, $LOV^{i,0}$ is changed to:

$$LOV^{i,0} = \max\left\{\frac{2}{3}\left(\sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^{\text{MD}}} + \sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^{PA}}\right)\right.$$
$$\left.\cdot \min_{l=\text{MA,PA},l\neq i}\left(1, \frac{\sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^i}}{\sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^l}}\right), \frac{2}{c^i_{\text{joint}}}, \sum_{k\in\{\text{room,joint,med}\}} \frac{1}{c_k^i}\right\}. \tag{3.20}$$

Moreover, the utilizations of the providers and MAs can be evaluated.

$$\rho_i = \widetilde{P}_2^i + \widetilde{P}_3^i + \widetilde{P}_5^i, \quad i = \text{MD, PA},$$
$$\rho_{\text{MA}} = \frac{\sum_{i=\text{MD,PA}}[2(\widetilde{P}_1^i + \widetilde{P}_2^i + \widetilde{P}_4^i + \widetilde{P}_5^i + \widetilde{P}_6^i) + \widetilde{P}_3^i]}{3}.$$

Following this procedure, the performance of the joint system with MA wrap-up can be analyzed. Table 3.9(a) presents the results compared with simulation under Markovian assumption and using the collected data. By changing the CVs, comparisons are carried out for MD patients and PA patients under general distributions where the results are shown in Table 3.10.

Using randomized data, the comparison results in Markovian cases are summarized in Table 3.9(b), which show that the average differences in throughput and LOV are still very small, within 3% and 1.5%, respectively. In addition, under non-exponential distributions, the average differences in LOV are 1.373% and 1.298% for MD and PA

Table 3.9: Accuracy of joint visit system with 3 MAs and MA wrap-up: Markov chain case

	Throughput(1/hr)		Length of visit(min)	
Provider team	MD	PA	MD	PA
Model	2.034	2.692	58.023	43.718
Simulation	1.981	2.628	58.320	42.951
ϵ^{exp}	2.674%	2.419%	0.509%	1.786%

(a) Collected data setting

	Throughput		Length of Visit	
Provider team	MD	PA	MD	PA
ϵ^{exp}	3.121%	2.111%	0.945%	1.542%

(b) Randomized data setting

Table 3.10: Accuracy of joint visit system with 3 MAs and MA wrap-up: general case

$(CV^{\text{eff}})^2$	$LOV^{\text{non-exp}}$	LOV^{sim}	$\epsilon^{\text{non-exp}}$
0.002	53.547	54.560	1.857%
0.161	54.262	54.830	1.037%
0.217	54.512	54.910	0.724%
0.288	54.830	55.900	1.914%
0.474	55.667	56.650	1.736%
0.477	55.680	56.700	1.799%
0.642	56.416	56.969	0.970%
0.729	56.806	57.310	0.879%
0.872	57.449	57.528	0.137%
0.940	57.755	57.710	0.078%

(a) MD team

$(CV^{\text{eff}})^2$	$LOV^{\text{non-exp}}$	LOV^{sim}	$\epsilon^{\text{non-exp}}$
0.029	39.622	40.100	1.191%
0.085	39.860	39.980	0.300%
0.180	40.259	40.390	0.325%
0.294	40,741	40.560	0.446%
0.309	40,804	40.830	0.065%
0.433	41.328	41.390	0.150%
0.504	41.626	42.200	1.361%
0.784	42.806	42.426	0.895%
0.824	42.977	42.439	1.269%
0.947	43.321	42.680	1.502%

(b) PA team

patients, respectively. Again, the introduced method leads to a close estimate of system performance.

3.4.3 Discussions

Monotonicity

Now, to improve the performance of the given system, it is necessary to understand the monotonicity properties with respect to service times. Here we consider the partial derivatives of throughput and utilization with respect to service rates. For the joint visit system with three MAs and provider wrap-up, the following monotonic properties are obtained.

Proposition 3.5 *In the joint visit system with three MAs and provider wrap-up, the system throughput is monotonically increasing with respect to service rate* c_k, $k \in \{\text{room}, \text{joint}, \textit{wrap}, \textit{med}\}$. *The utilizations of providers and MAs are monotonically increasing or decreasing with respect to service rate* c_k, $k \in \{\text{room}, \text{joint}, \text{wrap}, \text{med}\}$, *except MAs' utilization to* c_{joint}. *Such properties are summarized in Table 3.11:*

Table 3.11: Monotonicity with respect to service rates: joint system with 3 MAs and provider wrap-up

	c_{room}	c_{joint}	c_{wrap}	c_{med}
Throughput ($\frac{\partial TP}{\partial c_k}$)	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
Provider utilization ($\frac{\partial \rho_{\text{MD}}}{\partial c_k}$)	$\nearrow$	$\searrow$	$\searrow$	$\nearrow$
MA utilization ($\frac{\partial \rho_{\text{MA}}}{\partial c_k}$)	$\searrow$	$\times$	$\nearrow$	$\searrow$

where $\nearrow$ *and* $\searrow$ *indicates monotonically increasing and decreasing, respectively, and*

$\times$ *implies both can occur.*

Proof: See Appendix A. ∎

The monotonicity of throughput with respect to service rates are straightforward, i.e., higher rates or shorter service times lead to a higher throughput or larger patient volume. Similarly, the provider and MA utilizations with respect to c_{room} and c_{med} (which are MA rooming and MA visit rates, respectively) also coincide with intuition. When c_{room} and c_{med} are increased, less time in rooming and MA visit is taken so that the portion of provider's working time (joint visit and wrap-up) within a patient cycle time is increased and thus their utilization is higher, while the MAs spend less time so that their utilization decreases. On the contrary, when c_{wrap} is increased, provider takes less time to wrap up. Hence, the provider utilization is decreased and MA utilization is increased. Similarly, when c_{joint} is increased, providers spend less time in diagnosis, which results in a smaller utilization. This should in return lead to higher utilization of MA. However, the MA documentation time is also reduced, which should lead to lower utilization. Thus, the MA utilization with respect to joint visit rate is undetermined, dependent on the parameters of the system.

For example, consider the two scenarios shown in Figure 3.5. In the left part of the figure, $c_{\text{room}} = 0.147$, $c_{\text{joint}} = 0.224$, $c_{\text{wrap}} = 0.353$, and $c_{\text{med}} = 0.346$, while in the right part, all parameters are the same except $c_{\text{joint}} = 0.037$. In other words, the only difference between the two cases is the joint visit service time: 4.46 minutes and 26.77 minutes for the left and right cases, respectively. As illustrated in the right figure, when joint visit service rate is relatively small (i.e., taking relatively longer time), MA utilization is monotonically decreasing with respect to c_{joint}. In this case, as the joint visit time is significantly dominating and thus limiting the patient flow, its reduction frees

up the MA in some degree, but is not enough to increase much patient volume. When joint visit time is similar to other services, as shown in the left figure, MA utilization is monotonically increasing with respect to c_{joint}, i.e., increasing c_{joint} results in larger patient volume, and thus MAs are busier to serve more patients.

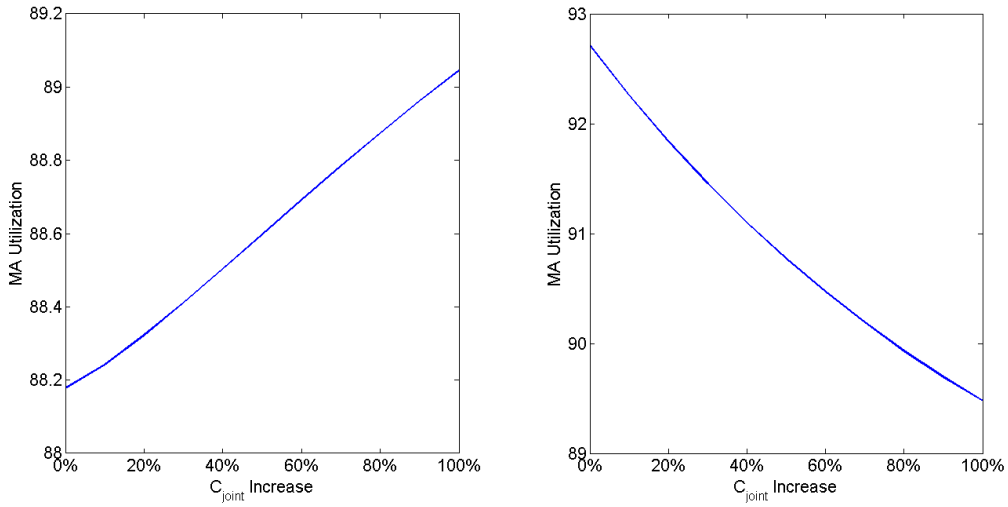

Figure 3.5: Monotonicity of ρ_{MA} with respect to c_{joint}: $c_{\text{room}} = 0.147$, $c_{\text{wrap}} = 0.353$, $c_{\text{med}} = 0.346$. c_{joint} starts from: 0.224 (left); 0.037 (right)

For the joint visit system with three MAs and MA wrap-up, the monotonicity properties are obtained with the same rationale explained above.

Proposition 3.6 *In the joint visit system with three MAs and MA wrap-up, the system throughput is monotonically increasing with respect to service rate c_k, $k \in \{\text{rooming}, \text{joint}, \text{MAvisit}\}$. The provider utilization is monotonically increasing with respect to c_{room} and c_{med} but decreasing with respect to c_{joint}. MA utilizations are always 100%.*

Proof: See Appendix A. ■

Wrap-up Scheme

So far in the introduced joint visit models, either provider or MA could wrap up the service. Then the question that naturally arises is: which wrap-up scheme is more efficient? To answer this question, we investigate the impact of the ratio between joint visit service time and total service time (i.e., sum of rooming time, joint visit time, wrap-up and MA visit time) on system throughput. Figure 3.6 illustrates the changes in throughput as a function of such a ratio, where the blue solid line indicates the provider wrap-up case and the red dash line refers to the MA wrap-up case. As one can see, provider wrap-up performs better when the ratio is low, and MA wrap-up is better when the ratio is high. The rationale behind this is that when joint visit spares longer time (i.e., higher ratio), provider will become the constraint of the system. Asking him/her to continue wrap-up will further slow down the process. This can also be justified by the right part of the curves, where higher ratios lead to a lower throughput. However, when joint visit is finished in short time, the majority of service times are on MA rooming and visit, which implies that the provider should be assigned more work, thus need to wrap up the visit. As shown in the figure, the crossover occurs roughly when the ratio is 0.5.

From the data collected at Dean Clinic, such ratios are 0.672 and 0.637 for MD team and PA team, respectively. Therefore, MA wrap-up is expected to lead to a higher throughput, and thus this conclusion has been recommended to Dean Health System.

Comparisons

Using the data collected at Dean East Clinic, we compare the performance of current system to the proposed joint visit system (considering both provider and MA wrap-up scheme) with an additional MA. The results are shown in Table 3.12 where $TP^{\text{non-exp}}$

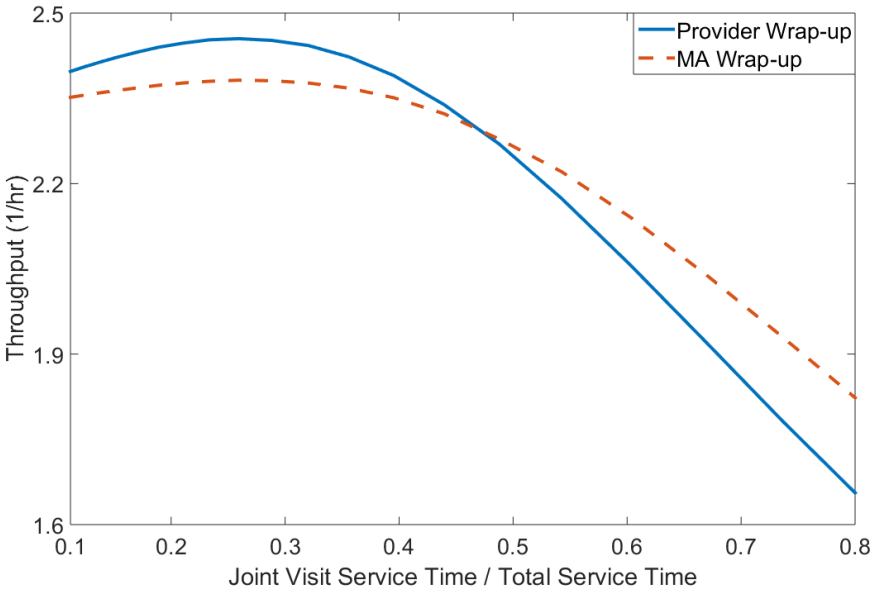

Figure 3.6: Comparison between provider wrap-up and MA wrap-up

is presented in (1/hr) and $LOV^{\text{non-exp}}$ in (min).

Table 3.12: Comparison between current system and joint visit systems

	$TP^{\text{non-exp}}$		$LOV^{\text{non-exp}}$		ρ_{Provider} (%)		ρ_{MA} (%)	
Provider team	MD	PA	MD	PA	MD	PA	MD	PA
Current	1.759	2.374	63.2	49.48	91.45%	88.04	34.21	40.76
Joint/Provider wrap-up	1.920	2.4973	60.83	46.80	96.84	96.77	78.08	
Joint/MA wrap-up	2.034	2.6915	54.28	40.46	90.74	88.59	96.14	

Summarizing the results, by adopting the joint visit service model with an additional MA, the MD patient throughput is increased by 9.15% and 15.64% for provider and MA wrap-ups, respectively, and the average length of visit is reduced by 3.75% and 14.13%,

respectively. Similar results are observed for PA patients. Such improvements are significant, which can help improve patient access substantially. In addition, workloads become much more balanced as shown in provider and MA's utilizations.

Clearly, by implementing the joint visit model with MA wrap-up, the potential benefit in throughput increase (about 2 patients per day per provider) can be significant in the long run compared with the price of an additional MA. Or equivalently, if the patient volume is kept constant, a substantial reduction in provider's workload could free them a significant amount of time from working after hours or on weekend, which again is valuable. Therefore, hiring an additional MA for joint visit with MA wrap-up has been recommended to Dean East Clinic, and the implementation is in progress.

Remark 3.2 The computation efforts using Procedures 3.1 and 3.2 are minimal. Typically the procedures converge within 5-7 iterations, which only takes a fraction of a second. Moreover, the proposed iteration method can be easily extended to more complex scenarios, such as more patient rooms and additional providers or MAs, without encountering the curse of dimensionality. For example, when adding one more provider (with two exam rooms) and an additional float MA, the state space will expand from 76 states to 1565 states if a direct Markov chain approach is used. Using the iterative approach, each time we still consider a system with 1 provider, 2 MAs and 2 exam rooms. Thus, the state number keeps at 9 only. The modification of the service rates of MA visit will then adjust the overestimation of the throughput. Therefore, such iteration method leads to an efficient calculation of system performance. □

3.4.4 Limitations and Extensions

Even though models introduced in this chapter aim to be as general as possible, healthcare delivery systems exhibit large variations in structure wise. Thus, some possible ways to extend the current models are introduced in this section.

First, the models discussed in this chapter assume that patients are always available for rooming, i.e., unlimited arrivals. This implies that patients always arrive on or before the scheduled appointment time. In practice, most patients do arrive earlier than their scheduled time. However, there still exist possibilities of arriving late, or even not showing up. To accommodate such scenarios, one possibility is to add arrival nodes in front of the rooming activity. The availability of these nodes can be characterized by the probability of on-time or early arrivals. This will lead to changes in balance equations, but the solution approach remains similar.

Secondly, the staff configuration is fixed to 2 providers with 3 float MAs per pod. As there is no standardized provider to MA ratio, this configuration may have a wide variation in practice. For example, a pod may consist of three providers with either three dedicated MAs or four shared float MAs. The former case is the same in structure wise as the current system with joint visit since each provider group can be analyzed separately. For the latter one, the iteration procedure needs to be modified such that the availability of the MA for one provider should be considered by taking into account whether the MA is serving patients of the other two providers. Thus, the iteration needs to circulate among these three providers.

Thirdly, current models can only handle the cases of two exam rooms per provider group. For different exam room settings, the state space needs to be redefined by taking into account all the feasible states. For instance, if one additional room is assigned

to each provider group, we need to consider all possible system states where MA and provider circulate around three exam rooms. After defining the new states and deriving the new balance equations, the remaining solution approach remains similar.

When applying to different system structures, the key issue is in redefining the system states. Once system states are identified, the remaining solution approach is relatively straightforward in implementing. Thus, the challenge in applicability is how to automatically search for all feasible states.

3.5 Conclusions

In this chapter, a Markov chain based method with extension to non-exponential case using an empirical formula is presented to study the joint visit service model in primary care clinics. The results show that the method can deliver effective and accurate estimation of system performance. Thus, it can be used as a quantitative tool for clinic managers to assess the impact of implementing joint visit models in clinics and provide guidelines in adopting the appropriate form of service model.

Chapter 4

Timeliness in Lung Cancer Diagnosis-to-treatment Process

4.1 Introduction

In this chapter, a system-theoretic method is introduced to analyze the diagnosis-to-treatment process for lung cancer patients who go through surgical resections. Section 4.2 describes the diagnosis-to-treatment process and formulates the problem. In Section 4.3, the complex care delivery process is decomposed into a collection of serial processes, and closed formulas are derived to evaulate the system performance. An approximation algorithm is developed in Section 4.4 to estimate the waiting-time performance; the probability to receive the surgery within a desired or given time period. In addition, to identify the bottleneck waiting time whose improvement will lead to the largest reduction in the overall waiting time, we present simple indicator measures that can be used in practice. Furthermore, the applicability of the proposed method is illustrated via a case study at Baptist Memorial Hospital in Section 4.5. Finally, conclusions and managerial insights for improvement are formulated in Section 4.6.

4.2 Process Description and System Modeling

4.2.1 Process Description

The lung cancer diagnosis-to-surgery process can be divided into the following five steps:

- Step 1: Chest X-ray and/or CT-scan.

  The diagnosis process of a lung cancer patient usually begins with an abnormal chest X-ray and/or CT-scan. Either test usually serves to initially identify the presence of a potentially cancerous lung lesion. The tests can be conducted exclusively or jointly.

- Step 2: Diagnostic biopsy.

  The diagnostic biopsy typically follows the abnormal chest X-ray and/or CT-scan to confirm a suspected lung cancer through tissue analysis. The major procedures in biopsy process include CT-guided biopsy and bronchoscopy, which again may be carried out exclusively or jointly.

- Step 3: Non-invasive staging.

  Multiple procedures may be carried out in the non-invasive staging step to evaluate the extent of spread of the lung cancer. The combinations of the procedures include CT-scan, position emission tomography (PET)/CT, brain imaging and bone scan. Either a single test, or a combination of any two or three of them, or even all the procedures can be conducted (which leads to a total of 15 options).

- Step 4: Invasive staging.

Invasive staging is carried out to confirm the extent of spread of lung cancer. Such tests include endobronchial ultrasound (EBUS) and mediastinoscopy (MED) where patients may receive either one or both of them. Invasive staging tests can be used concurrently for initial histologic diagnosis and staging in the optimal, parsimonious approach to care delivery.

- Step 5: Surgery.

  Finally, patients with early disease stage receive surgery as the last step. Prior to surgery, patients may need to receive the required medical clearance, i.e., consultation with surgeon or cardiac specialists to confirm the patient is healthy enough to endure surgery. Patients may receive consultation in different sequence, or in some cases skip one or both visits.

The standard lung cancer diagnosis-to-surgery procedure is defined to follow steps 1-5 in a sequential order. However, even though there exists a standard procedure, its actual execution depends on a number of issues and has a high variation. During the pathway along the diagnosis-to-surgery process, patients may skip or repeat certain steps, or follow a reversed sequence (see Figure 4.7) such as going back to prior steps for missing tests or retesting.

4.2.2 System Modeling

As shown in Figure 4.7, the diagnosis-to-surgery process can be viewed as a complex network with multiple splits, merges, and parallel lanes. Evaluating the performance of such a network is the goal of this chapter. To do so, we first define the waiting time between Steps i and j as $w_{i,j}$, $i = 1, \ldots, 4$, $j \neq i$, $j = 1, \ldots, 5$. It is assumed that $w_{i,j}$ follows a general (i.e., arbitrary) distribution with mean $\tau_{i,j}$ and CV $cv_{i,j}$. In addition,

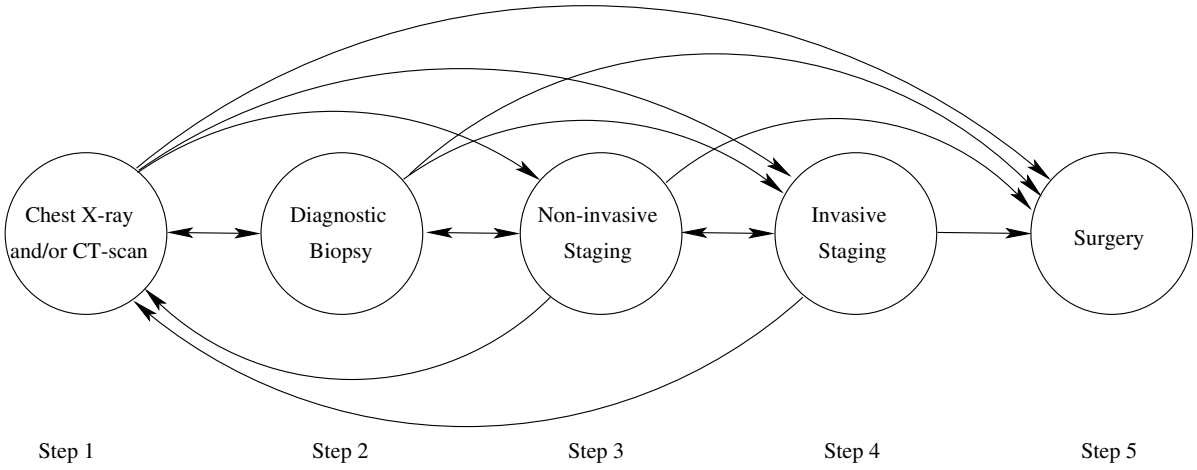

Figure 4.7: Lung cancer diagnosis-to-surgery process

denote the probability of a patient going from Step i to Step j as $\alpha_{i,j}$. Note that we do not consider the diagnosis or test procedure times in the model. While the procedure times of the diagnosis tests are typically short (minutes or hours), substantial waiting times occur in days or even months between the tests. Thus the test times are assumed negligible and we mainly focus on the waiting times. In addition, we assume the waiting times are independent to the past history of the patient's pathway along the process .

Remark 4.1 The independence assumption implies that the routing probabilities for different pathways and the waiting times are independent. In practice, such probabilities and waiting times may be correlated to patient factors such as willingness to consent and clinical condition ("performance status"). Since the focus of this chapter is on the average performance for general patients, these assumptions can be used as a starting point for analysis. In future work, we plan to address the scenarios related to patient specific pathways and waiting time correlations for specific patient groups.

Denote the number of waiting times as M. Within the diagnosis-to-surgery process, there are a total of $M = 16$ waiting times. Introduce vector $\mathcal{W}$ to denote these waiting

times:

$$\mathcal{W} = [w_{1,2}, w_{1,3}, w_{1,4}, w_{1,5}, w_{2,1}, w_{2,3}, w_{2,4}, w_{2,5},$$

$$w_{3,1}, w_{3,2}, w_{3,4}, w_{3,5}, w_{4,1}, w_{4,2}, w_{4,3}, w_{4,5}].$$

Let t denote the waiting time of the entire diagnosis-to-surgery process. Here t is a random variable with mean T and coefficient of variation CV, which are functions of all $\tau_{i,j}$'s, $cv_{i,j}$'s, and $\alpha_{i,j}$'s, i.e.,

$$T = E(t) = f_t(\mathcal{T}, \mathcal{V}, \mathcal{A}), \tag{4.1}$$

$$CV = \frac{\sqrt{Var(t)}}{T} = f_v(\mathcal{T}, \mathcal{V}, \mathcal{A}), \tag{4.2}$$

where

$$\mathcal{T} = [\tau_{1,2}, \tau_{1,3}, \tau_{1,4}, \tau_{1,5}, \tau_{2,1}, \tau_{2,3}, \tau_{2,4}, \tau_{2,5},$$

$$\tau_{3,1}, \tau_{3,2}, \tau_{3,3}, \tau_{3,5}, \tau_{4,1}, \tau_{4,2}, \tau_{4,3}, \tau_{4,5}],$$

$$\mathcal{V} = [cv_{1,2}, cv_{1,3}, cv_{1,4}, cv_{1,5}, cv_{2,1}, cv_{2,3}, cv_{2,4}, cv_{2,5},$$

$$cv_{3,1}, cv_{3,2}, cv_{3,3}, cv_{3,5}, cv_{4,1}, cv_{4,2}, cv_{4,3}, cv_{4,5}],$$

$$\mathcal{A} = [\alpha_{1,2}, \alpha_{1,3}, \alpha_{1,4}, \alpha_{1,5}, \alpha_{2,1}, \alpha_{2,3}, \alpha_{2,4}, \alpha_{2,5},$$

$$\alpha_{3,1}, \alpha_{3,2}, \alpha_{3,3}, \alpha_{3,5}, \alpha_{4,1}, \alpha_{4,2}, \alpha_{4,3}, \alpha_{4,5}].$$

In addition, the waiting-time-performance, WTP, i.e., the probability to finish the diagnosis-to-surgery process within a desired time interval ω_d, is defined as

$$WTP = \text{Prob}\{t \leq \omega_d\} = f_w(\mathcal{T}, \mathcal{V}, \mathcal{A}, \omega_d). \tag{4.3}$$

which again is a function of all these parameters.

The problem to be addressed in this chapter is to develop a method to evaluate and improve T, CV, and WTP.

4.3 Mean and Coefficient of Variation of Waiting Time

4.3.1 Decomposition

The diagnosis-to-surgery process is a complex process with multiple splits, merges, and random waiting times each following an unknown distribution type. All these factors complicate the study of such a network, particularly making the direct evaluation of the process variability almost impossible. However, considering that all patients are independent, and each patient can only take one specific pathway, we can decompose such a complex network into a collection of serial processes. Thus in this framework, each serial process represents a pathway a patient may follow. For example, a patient who receives a chest X-ray (Step 1) may need a biopsy test (Step 2) afterwards with probability $\alpha_{1,2}$, then may require invasive staging (Step 4) with probability $\alpha_{2,4}$, and finally go through surgery (Step 5) with probability $\alpha_{4,5}$. In this sense, this patient's pathway can be described by a serial process (Step 1 $\rightarrow$ Step 2 $\rightarrow$ Step 4 $\rightarrow$ Step 5) with probability $\alpha_{1,2}\alpha_{2,4}\alpha_{4,5}$. Examples of ten pathways are illustrated in Figure 4.8.

Using such a decomposition method, the complex process is represented by multiple serial processes, where each one is a combination of various step options. Since there exist various sequences of tests in the process, and the same tests may be taken multiple times, in theory, there could be infinite number of serial processes. However, given some physical constraints (e.g., some procedures may not be taken multiple times), and by ignoring the pathways that have negligible small probabilities, we can obtain a finite number of serial processes for analysis, and we denote this number as K. These processes

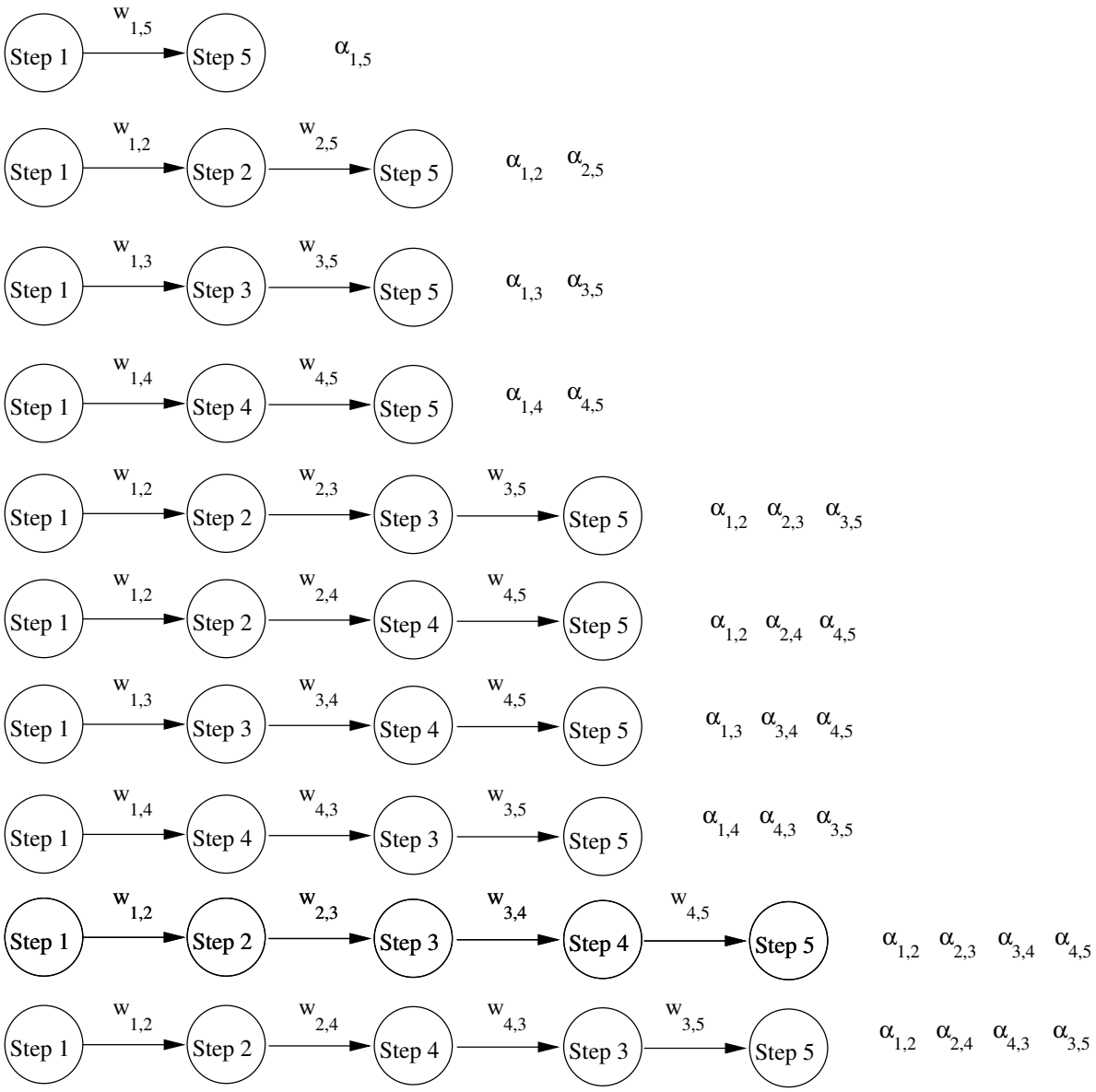

Figure 4.8: Ten pathway examples

are denoted as 1, 2, through K.

Introduce vector $\mathcal{S}_i$, $i = 1, \ldots, K$, whose element $\theta_{i,j}$, $j = 1, \ldots, M$, represents the number of times the waiting time j has been gone through in the i-th pathway. Here, the waiting time j's are in the same sequence as in vector $\mathcal{W}$. For example, $\theta_{4,2} = 1$ implies that the waiting time $w_{1,3}$ (from Step 1 to Step 3) has been taken once in the 4-th serial process, while $\theta_{4,2} = 2$ indicates that such a waiting has been taken twice in this pathway. Then,

$$\mathcal{S}_i = [\theta_{i,1}, \theta_{i,2}, \ldots, \theta_{i,M}], \quad i = 1, \ldots, K.$$

For each serial process, by multiplying the routing probabilities at each step, we can obtain the probability of a patient taking this pathway by

$$p_i = \prod_{k=1}^{M} (\mathcal{A}_k)^{\theta_{i,k}}, \quad i = 1, \ldots, K,$$

where $\mathcal{A}_k$ is the k-th element in routing probability vector $\mathcal{A}$.

4.3.2 Performance Measure

For each serial process, the mean waiting time is evaluated by summing up all the average waiting times included in the pathway.

$$T_i = \mathcal{S}_i \cdot \mathcal{T}', \quad i = 1, \ldots, K, \tag{4.4}$$

where $\mathcal{T}'$ indicates the transpose of vector $\mathcal{T}$.

Given the performance of each serial process, we obtain the performance of the entire diagnosis-to-surgery process by calculating the weighted average of the serial processes. Note that since we only consider a finite number of serial processes, the total probability of K serial processes will be smaller than one. Hence, normalization needs to be carried

out, i.e., divide by the total probability.

$$T = \frac{\sum_{i=1}^{K} p_i T_i}{\sum_{i=1}^{K} p_i}. \tag{4.5}$$

Unlike the mean, generally it is impossible to evaulate the variance even for a serial process. However, as waiting times are assumed to be independent (i.e., covariance = 0), the variance of a serial process can be obtained by the sum of variances of individual waiting times. Thus, given the variance of each serial pathway, the variability of the whole diagnosis-to-surgery process can be evaluated.

Proposition 4.1 *The variance of waiting time can be evaluated as*

$$CV = \frac{\sqrt{\frac{\sum_{i=1}^{K} p_i Var_i}{\sum_{i=1}^{K} p_i}}}{T}, \tag{4.6}$$

where T is evaluated from (4.5), and Var_i is the variance of waiting time for the i-th pathway and can be calculated by

$$Var_i = \sum_{j=1}^{M} \theta_{i,j} \mathcal{T}_j^2 (\mathcal{V}_j^2 + 1) - \Big(\sum_{j=1}^{M} \theta_{i,j} \mathcal{T}_j \Big)^2, \tag{4.7}$$

and $\mathcal{T}_j$ and $\mathcal{V}_j$ are the j-th elements of vectors $\mathcal{T}$ and $\mathcal{V}$, respectively.

Proof: See Appendix B. ∎

It can be shown that both T and CV have monotonic properties with respect to its arguments, i.e.,

Proposition 4.2 *The mean and the coefficient of variation of waiting time, T and CV, are monotonically increasing with respect to $\tau_{i,j}$ and $cv_{i,j}$, respectively.*

Proof: See Appendix B. ∎

Therefore, to reduce waiting time and its variability, efforts should be focused on decreasing $\tau_{i,j}$ and $cv_{i,j}$. However, decreasing which $\tau_{i,j}$ or which $cv_{i,j}$ will lead to the largest reduction in T and CV, respectively, is related to identifying the bottlenecks, which is introduced in the next section.

4.3.3 Bottleneck Analysis

Bottleneck analysis is viewed as the most effective way to improve system performance in production systems research [93]. Such an effort is also useful for healthcare delivery systems (e.g., [53, 94, 95, 96, 97]) where the bottlenecks are also commonly referred to as constraints or barriers. A bottleneck factor (a procedure or resource) is the one that impedes the system performance in the strongest manner, i.e., whose improvement will lead to the largest improvement in the overall system performance.

Mean waiting time bottleneck

By definition, bottlenecks are identified by evaluating the partial derivatives with respect to an argument (see for instance, [94, 93, 96, 97]). However, when considering average waiting time, using a common incremental change (i.e., using partial derivative) may ignore the fact that efforts to reduce waiting times can vary substantially in scale if the average waiting times are significantly different. Therefore, here we introduce the mean waiting time bottleneck as follows:

Definition 4.1 *Waiting time mean $\tau_{i,j}$ is the mean waiting time bottleneck (BN-τ) if*

$$T(\tau_{i,j} - \delta_\tau \tau_{i,j}) < T(\tau_{k,l} - \delta_\tau \tau_{k,l}), \quad \forall \{k,l\} \neq \{i,j\},$$

where $T(\tau_{i,j} - \delta_\tau \tau_{i,j})$ represents the resulting waiting time when $\tau_{i,j}$ is reduced by proportion δ_τ, $\delta_\tau \ll 1$.

Although the bottlenecks can be discovered by directly applying the definitions (manually changing each waiting time and comparing with the original case), this approach involves many computation efforts. In addition, this does not explicitly show why a specific waiting time is the bottleneck. Therefore, we seek some indirect measurements that do not involve much computation efforts and also can easily explain the rationale of bottleneck waiting time.

To identify the BN-τ, we obtain

Proposition 4.3 *Denote the k-th element in vector* $\mathcal{T}$ *as* $\mathcal{T}_k$*. Then*

$$T(\mathcal{T}_k - \delta\mathcal{T}_k) > T(\mathcal{T}_j - \delta\mathcal{T}_j), \text{ if and only if}$$

$$\sum_{i=1}^{K} p_i\theta_{i,k}\mathcal{T}_k < \sum_{i=1}^{K} p_i\theta_{i,j}\mathcal{T}_j, \quad j \neq k.$$

Proof: See Appendix B. ∎

Using this result, define

$$I_{\tau,k} = \sum_{i=1}^{K} p_i\theta_{i,k}\mathcal{T}_k. \tag{4.8}$$

By finding the largest $I_{\tau,k}$, we can identify the most impeding waiting time. Thus, $I_{\tau,k}$ can be used as an indicator for BN-τ identification.

BN Indicator 1: The mean time bottleneck (BN-τ) is the waiting time that has the largest value of $I_{\tau,k}$, $\forall k = 1, \ldots, M$.

This indicator implies that the waiting point many patients go through and at the same time has long average wait time is the one whose reduction will lead to the largest decrease in overall waiting time. The rationale behind this is that the weighted time (average waiting time multiplied by routing probability) indicates the priority of a waiting time's impact on the whole process.

Waiting time variability bottleneck

For healthcare systems, since CVs are generally less than 1 (i.e., the longer the service has been carried out, the higher the probability the service will be finished), we can safely assume the CVs are similar in scale. Thus the waiting time variability bottleneck can be defined as:

Definition 4.2 *Waiting time coefficient of variation $cv_{i,j}$ is the variability bottleneck (BN-cv) if*

$$\frac{\partial CV}{\partial cv_{i,j}} > \frac{\partial CV}{\partial cv_{k,l}}, \quad \forall \{k,l\} \neq \{i,j\}.$$

To identify the variability bottleneck (BN-*cv*), we obtain

Proposition 4.4 *Denote the k-th element in vector $\mathcal{V}$ as $\mathcal{V}_k$, then*

$$\frac{\partial CV}{\partial \mathcal{V}_k} > \frac{\partial CV}{\partial \mathcal{V}_j}, \text{ if and only if}$$

$$\sum_{i=1}^{K} p_i \theta_{i,k} \mathcal{T}_k^2 \mathcal{V}_k > \sum_{i=1}^{K} p_i \theta_{i,j} \mathcal{T}_j^2 \mathcal{V}_j, \quad j \neq k.$$

Proof: See Appendix B. ∎

Then, define

$$I_{cv,k} = \sum_{i=1}^{K} p_i \theta_{i,k} \mathcal{T}_k^2 \mathcal{V}_k. \tag{4.9}$$

We obtain the most impeding waiting time in terms of variability by finding the largest $I_{cv,k}$:

BN Indicator 2: The variability bottleneck (BN-*cv*) is the waiting time that has the largest value of $I_{cv,k}$, $\forall k = 1, \ldots, M$.

As one can see, such an indicator depends on both the weighted probability, the mean waiting time, and the coefficient of variation of the waiting time. If many patients

go through a waiting point that has a large mean and variability, then this could be the most impeding one that efforts should be focused on.

4.4 Waiting-time Performance

4.4.1 Approximation Formula

In many cases, knowing the first two moments, i.e., the mean and variance (or coefficient of variation), is not enough to fully understand a process. Often times, more detailed information such as a complete distribution is needed. Specifically, the probability to finish the process within a desired or given time interval, referred to as the waiting-time performance or WTP, characterizes such information. However, evaluating WTP for the lung cancer diagnosis-to-surgery process is not straight forward as waiting times follow an arbitrary distribution.

To solve such an issue, an approximation method using Gamma distribution is proposed. First, we hypothesize that the WTP is practically independent of distribution type, but mainly depends on the first two moments, i.e., mean and CV. Under this hypothesis, the process with general distribution can be approximated by a process following Gamma distribution with the same mean and variance. Thus, using Gamma distributions to represent each waiting time with given mean and CV, we can evaluate the WTP.

From equation (4.3),

$$WTP = \text{Prob}\{t \leq \omega_d\} = \frac{\sum_{i=1}^{K} \text{Prob}\{t_i \leq \omega_d\} \cdot p_i}{\sum_{i=1}^{K} p_i},$$

where t_i is the waiting time of the i-th pathway, $i = 1, \ldots, K$.

Since Prob$\{t_i \leq \omega_d\}$ is the only unknown term in the equation, we then derive formulas to evaluate this term. Assuming that each waiting time follows Gamma distribution, Prob$\{t_i \leq \omega_d\}$ is equivalent to evaluating $G(\omega_d)$, the CDF of the sum of independently distributed Gamma variables which can be expressed as a single Gamma series.

Proposition 4.5 *The WTP to finish diagnosis-to-surgery process within ω_d can be calculated as*

$$WTP = \frac{\sum_{i=1}^{K} p_i \prod_{j=1}^{\sum_{l=1}^{M}\theta_{i,l}} \left(\frac{\beta_{min}}{\beta_j}\right)^{\alpha_j}}{\sum_{i=1}^{K} p_i} \cdot \sum_{k=0}^{\infty} \frac{\delta_k \gamma(\rho+k, \omega_d/\beta_{min})}{\Gamma(\rho+k)}, \tag{4.10}$$

where

$$\begin{aligned}
\alpha_i &= \frac{\mu_i^2}{\sigma_i^2}, \quad \beta_i = \frac{\sigma_i^2}{\mu_i}, \\
\beta_{min} &= \min(\beta_i), \quad i = 1, \ldots, n, \\
\rho &= \sum_{i=1}^{n} \alpha_i, \quad \delta_0 = 1, \\
\delta_{k+1} &= \frac{1}{k+1} \sum_{i=1}^{k+1} i\nu_i \delta_{k+1-i}, \quad k = 1, 2, \ldots,
\end{aligned} \tag{4.11}$$

$$\begin{aligned}
\nu_k &= \sum_{i=1}^{n} \alpha_i (1 - \beta_{min}/\beta_i)^k, \quad k = 1, 2, \ldots, \\
\gamma(a, x) &= \int_0^x y^{a-1} e^{-y} dy, \\
\Gamma(a) &= \lim_{x \to \infty} \gamma(a, x).
\end{aligned}$$

Proof: See Appendix B. ∎

4.4.2 Validation

To validate the approximation formula, we carry out dozens of simulation experiments using randomly generated mean (between 10 and 90) and CV (between 0.1 and 1). Validation is carried out by comparing the results obtained from Proposition 4.5 and simulation models. For the simulation models, the distributions of waiting times are chosen from either Weibull or Lognormal because both have two parameters which provides freedom in selecting the coefficient of variation. In addition, a mixed distribution where each waiting time randomly follows Lognormal, Weibull or Gamma is also generated.

Table 4.13 presents the average, minimum, and maximum differences compared with the simulation results. Four examples are also shown in Figure 4.9 where the blue line represents the WTP obtained through Proposition 4.5 and the red line refers to simulation results using mixed distribution of Lognormal, Weibull and Gamma. The results indicate that WTP is practically independent of distribution type, and the results from analytical formula closely match with simulations under random mean and CV settings. Similar results are observed under other parameter settings with randomized routing probabilities as well. Thus we conclude that the developed model can be used to evaluate the waiting-time performance in general settings.

4.4.3 WTP Bottleneck

First, through extensive numerical studies, we observe that reducing the mean and coefficient of variation of any of the waiting times leads to an improvement of WTP. As illustrated from the examples in Figures 4.10 and 4.11, when ω_d is not sufficiently small, reducing $\tau_{i,j}$ and $cv_{i,j}$ always leads to an increase in WTP. All other $w_{i,j}$'s not shown in the examples also exhibit the similar monotonic behavior.

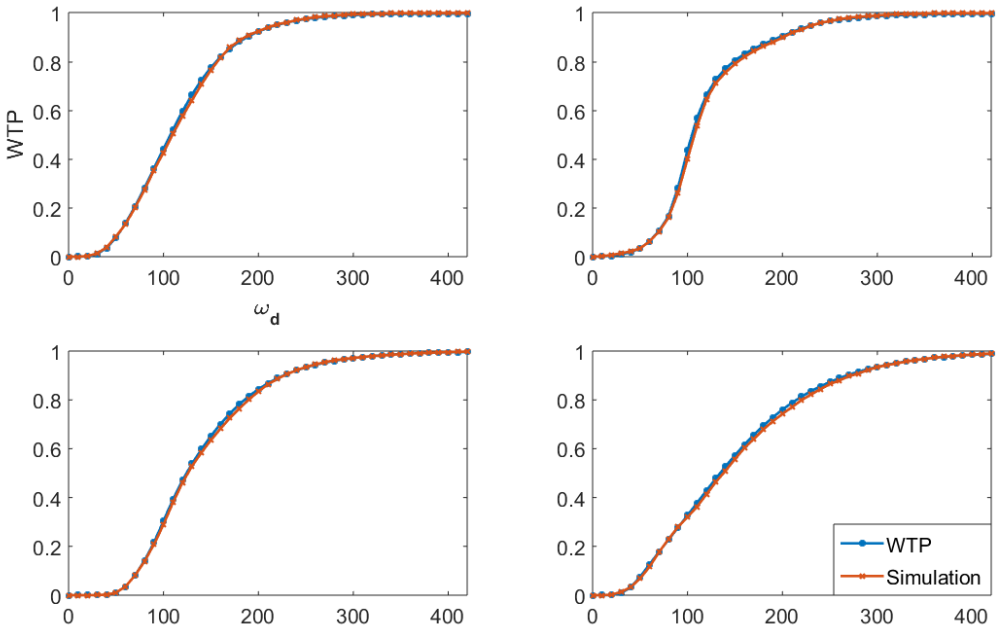

Figure 4.9: Comparison examples: Case 1: Mean=116; Case 2: Mean=121; Case 3: Mean=132, Case 4: Mean=153

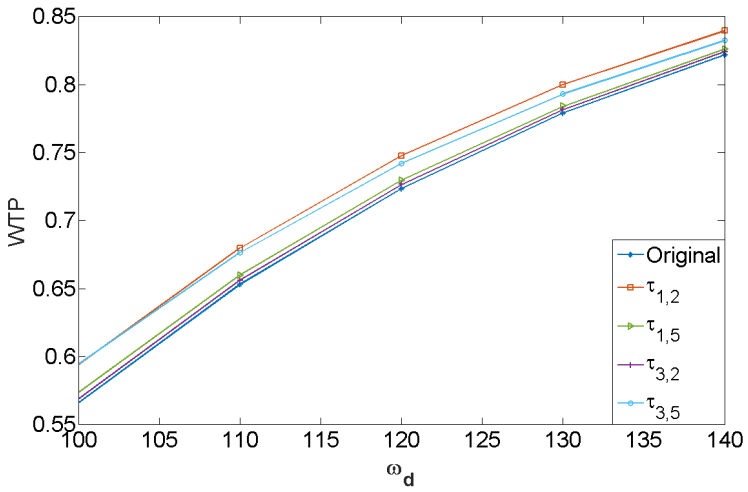

Figure 4.10: WTP monotoncity with respect to mean waiting time

Table 4.13: WTP comparison with simulation results

ω_d	Ave. diff.	Min. diff.	Max. diff.
10	0.000384	8.086E-09	0.001266
20	0.001042	3.138E-05	0.003176
30	0.001920	0.000147	0.004865
40	0.001754	1.296E-05	0.003104
50	0.002133	1.279E-06	0.006167
60	0.003745	0.000106	0.009251
70	0.005936	3.500E-05	0.019692
80	0.009582	0.001052	0.025943
90	0.009182	0.000161	0.023920
100	0.011788	0.002757	0.034184
110	0.013908	0.003105	0.033320
120	0.012810	0.003873	0.019612
130	0.014106	0.000894	0.022292
140	0.014703	0.003830	0.021533
150	0.013647	0.007073	0.019511
160	0.013718	0.002693	0.024230
170	0.014501	0.003642	0.025586
180	0.012083	0.003767	0.019501
190	0.010130	0.003458	0.017278
200	0.007668	0.001058	0.016816
210	0.005142	0.001664	0.016756
220	0.003356	9.545E-05	0.013780
230	0.002687	0.000223	0.013215

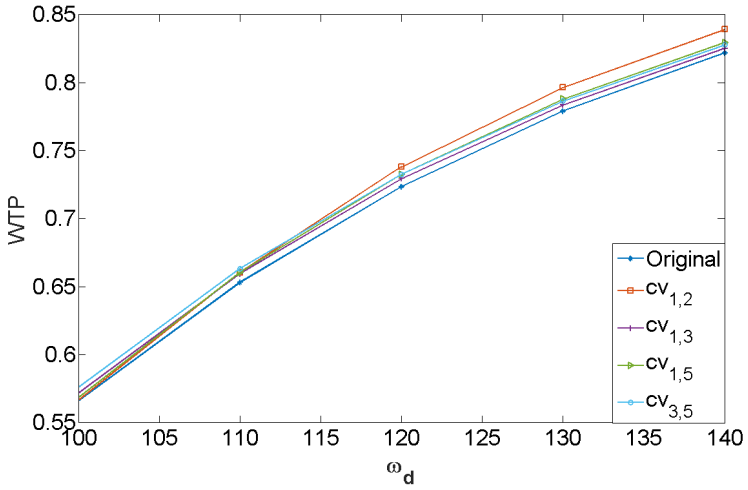

Figure 4.11: WTP monotoncity with respect to CV of waiting time

Remark 4.2 Note that the monotonicity with respect to $cv_{i,j}$ may not always hold when ω_d is very small. For small ω_d, majority of patients will not be able to finish the process within ω_d. Thus the ones that finish within ω_d will be more like outliers, which tends to appear more in larger variability settings. Thus reducing $cv_{i,j}$ may actually lead to less outliers and eventually decrease WTP. However, the probability to finish the whole diagnosis process within a short amount of time, such as 0-30 days (when on average it takes 95 days), is of less interest to practitioners. It would be more practical to provide the probability to finish within the average waiting time, or to avoid the worst case, give estimates on large ω_d values. For those ranges, $cv_{i,j}$ exhibits monotonicity behavior.

Using the monotonic property, the bottleneck waiting time with respect to WTP can be defined as:

Definition 4.3 *Waiting time $w_{i,j}$ is the waiting-time performance bottleneck (BN-wtp)*

if $\forall\{k,l\} \neq \{i,j\}$,

$$WTP(\tau_{i,j} - \delta_\tau \tau_{i,j}, cv_{i,j} - \delta_{cv} cv_{i,j})$$
$$> WTP(\tau_{k,l} - \delta_\tau \tau_{k,l}, cv_{i,j} - \delta_{cv} cv_{i,j}),$$

where $WTP(\tau_{i,j} - \delta_\tau \tau_{i,j}, cv_{i,j} - \delta_{cv} cv_{i,j})$ *represents the resulting waiting-time performance when both* $\tau_{i,j}$ *and* $cv_{i,j}$ *are reduced by proportions* δ_τ *and* δ_{cv}*, respectively, and* $\delta_\tau \ll 1$, $\delta_{cv} \ll 1$.

Unfortunately, a close-form formula to identify the BN-*wtp* is not available due to computational complexity, as even the evaluation of WTP is an approximation. However, by individually modifying each $w_{i,j}$ and comparing with the original case, the most impeding waiting time can be identified whose improvement will lead to the largest improvement in WTP.

4.5 Case Study

To illustrate the applicability of the developed analytical model, data collected in Baptist Memorial Health Care system is used. A total of 614 patients underwent attempted surgical resection for a suspected lung cancer during the time period of January 1, 2009 to June 30, 2013 within the Baptist Memorial Health Care system in Memphis, TN and Southaven, MS.

4.5.1 Mean Waiting Time and Variability

The care delivery process encompassing the entire time period from initial detection to surgery is segmented into a serial process of five steps. The minimum, maximum, mean

and CV of the waiting time between any two steps are listed in Table 4.14 and the routing probabilities are listed in Table 4.15.

By limiting to at most one re-test at each step, we obtain 379 routes that represent a total of 95.63% probability. In other words, this excludes 4.37% of routes that are assumed highly implausible in real scenario. Among the 379 routes, the route with the smallest probability (3.7E-10) is Steps 1-4-3-2-1-4-3-2-5, where all diagnostic tests are taken twice. This also indicates that the non-included routes have even smaller, thus negligible, probabilities. Therefore, in this study, we select $K = 379$, and $\sum_{i=1}^{K} p_i = 0.9563$.

Using these data and formula (4.5), the average waiting time is calculated as 95.96 days and the variance is evaluated as 1802.2 days which leads to $CV = 0.4424$.

4.5.2 Bottleneck and Improvement Analysis

To identify the bottlenecks, we first consider the mean waiting time bottleneck BN-τ. Using BN Indicator 1, we observe that the waiting times between Steps 1 and 2 and between Steps 3 and 5 are the main constraints. As shown in Table 4.16, $I_{\tau,k}$ is similar for both cases (27.59 and 25.98) which are substantially higher than all others. Therefore, we view both waiting times as BN-τ's.

In terms of BN-*cv*, again from Table 4.16, the variability of waiting time between Steps 1 and 2 results in the largest $I_{cv,k}$, and the waiting time between Steps 3 and 5 gives the second largest. Therefore, both waiting times are identified as the BN-*cv*'s as well. Hence, the bottleneck results imply that improvement efforts should be focused to reduce both the mean and variation of these two waiting times.

Next, to improve the identified bottleneck waiting times, we further consider the detailed test procedures for each bottlenecks. First, regarding the waiting time between

Table 4.14: Waiting time data

	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	$w_{1,5}$	$w_{2,1}$	$w_{2,3}$	$w_{2,4}$	$w_{2,5}$	$w_{3,1}$	$w_{3,2}$	$w_{3,4}$	$w_{3,5}$	$w_{4,1}$	$w_{4,3}$	$w_{4,5}$
Min	11	13	17	43	12	7	13	26	8	8	11	25	15	6	9
Max	114	98	84	189	64	19	50	68	89	57	57	76	45	85	78
Mean	40	35	42	93	29	11	30	42	37	20	27	44	30	33	25
CV	0.69	0.58	0.43	0.42	0.63	0.33	0.36	0.28	0.72	0.58	0.52	0.31	0.71	0.91	1.08

Table 4.15: Routing probabilities

$\alpha_{1,2}$	$\alpha_{1,3}$	$\alpha_{1,4}$	$\alpha_{1,5}$	$\alpha_{2,1}$	$\alpha_{2,3}$	$\alpha_{2,4}$	$\alpha_{2,5}$	$\alpha_{3,1}$	$\alpha_{3,2}$	$\alpha_{3,4}$	$\alpha_{3,5}$	$\alpha_{4,1}$	$\alpha_{4,3}$	$\alpha_{4,5}$
0.689	0.224	0.008	0.079	0.047	0.688	0.120	0.145	0.038	0.215	0.071	0.676	0.023	0.136	0.841

Table 4.16: BN-τ and BN-cv identification

	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	$w_{1,5}$	$w_{2,1}$	$w_{2,3}$	$w_{2,4}$	$w_{2,5}$	$w_{3,1}$	$w_{3,2}$	$w_{3,4}$	$w_{3,5}$	$w_{4,1}$	$w_{4,3}$	$w_{4,5}$
$I_{\tau,k}$	27.59	7.97	0.35	7.92	0.86	6.05	3.17	5.49	0.81	2.53	1.58	25.98	0.08	0.65	3.75
$I_{cv,k}$	0.19	0.04	0.00	0.08	0.00	0.01	0.01	0.02	0.01	0.01	0.01	0.09	0.00	0.00	0.03

Steps 1 and 2, Table 4.17 summarizes the percentage of patient volumes taking different procedures for each step and the corresponding waiting times. Using the similar criterion of BN Indicator 1, we discover that the waiting time between CT-scan and CT-guided biopsy is the main constraint, which takes an average of 86 days for almost one third of patients. Thus, reducing this waiting time can lead to the largest improvement compared with reducing all other waiting times.

Regarding the bottleneck waiting time between Step 3 and Step 5, different medical clearance consultations take place before the surgery where Table 4.18 summarizes the data characterizing these waiting times. We can observe that more than half of the patients require both cardiac and surgical consultations which brings additional time lags. Moreover, for the patients who require both cardiac and surgical consultations, it is found that if surgical consultation is taken first, it helps reduce the waiting time for cardiac consultation substantially. Thus, if both consultations are required, it is recommended that surgical consultation should be scheduled first.

4.5.3 Waiting-time Performance

Prior to evaluating the waiting-time performance, we first show that WTP is practically independent of the distribution type, and thus Gamma distribution can be used for approximation. To do so, Lognormal, Weibull, and mixed distributions are chosen to characterize the waiting times, and the results of these distributions are compared with the WTPs obtained from Gamma approximation. As shown in Figure 4.12 and Table 4.19, the differences are minimal which suggests that the Gamma approximation provides a high fidelity of WTP estimation.

In addition, to investigate the impact of variability on the performance, CV is varied

Table 4.17: BN-τ analysis in Steps 1 and 2 waiting time

Step 1	Step 2	Percentage	Avg. waiting	Weighted waiting
CT-scan	CT-guided biopsy	32.8%	86 days	28.21
CT-scan	Bronchoscopy	5.6%	18 days	1.01
CT-scan	CT-guided biopsy & bronchoscopy	3.2%	22 days	0.70
Chest X-ray & CT-scan	CT-guided biopsy	42%	19 days	7.98
Chest X-ray & CT-scan	Bronchoscopy	12%	17 days	2.04
Chest X-ray & CT-scan	CT-guided biopsy & bronchoscopy	4.4%	24 days	1.06

Table 4.18: BN-τ analysis in Steps 3 and 5 waiting time

Step 3	Waiting	Consult	Waiting	Consult	Waiting	Step 5	Percentage	Weighted waiting
Non-invasive staging	30 days	Surgical	5 days	Cardiac	18 days	Surgery	32.3%	17.1
	26 days	Cardiac	22 days	Surgical	12 days		30.1%	18.1
	24 days	Surgical	-	-	12 days		19.5%	7.0
	18 days	Cardiac	-	-	23 days		15.7%	6.4
	-	-	-	-	52 days		2.4%	1.2

from 0.1 to 0.9 in increments of 0.1 where all waiting time CVs are set identically. As one can see from Figure 4.13, the Gamma approximation results in good accuracy regardless of CV for all distribution types. However, while the differences in waiting time distributions are minimal under various CV settings, the distribution shapes are affected by the CVs. The WTP decreases as CV increases, which is consistent with intuition that system deteriorates when more variability is involved.

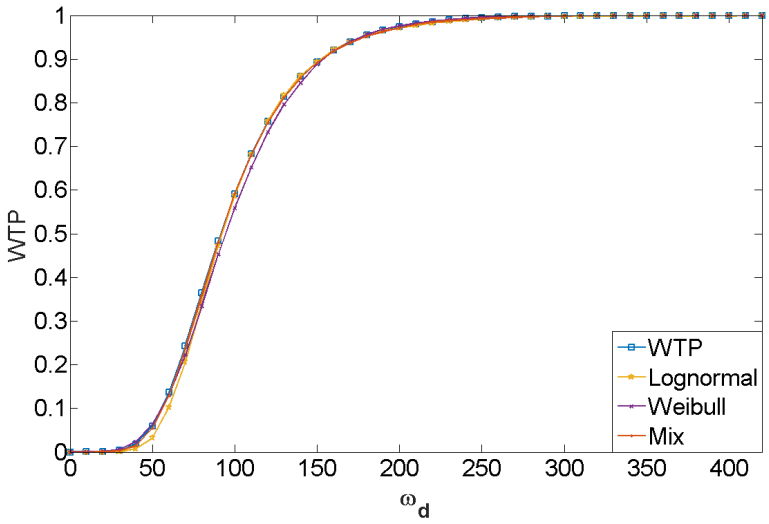

Figure 4.12: Comparison of Lognormal, Weibull, and mix distribution with Gamma approximation

Table 4.19: WTP comparison with Gamma approximation

Difference	Lognormal	Weibull	Mixed Distribution
Average	0.004459	0.005624	0.004261
Maximum	0.035185	0.031881	0.018007
Minimum	1.05E-05	7.08E-06	1.09E-04

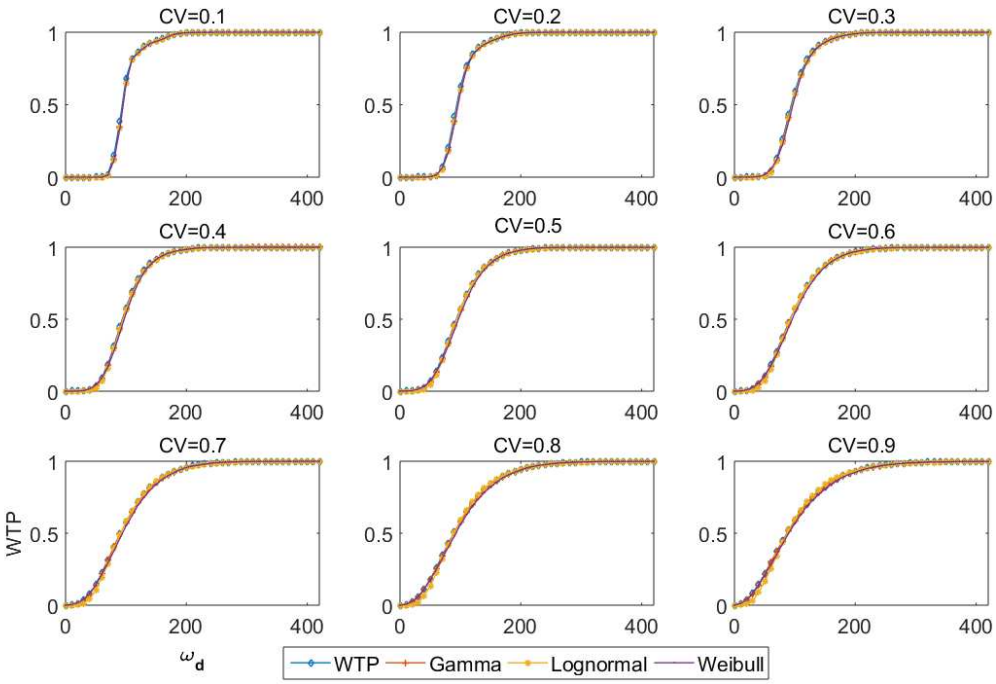

Figure 4.13: Various distributions under varying CV

Table 4.20: BN-*wtp* identification

ω_d	$w_{1,2}$	$w_{1,3}$	$w_{1,4}$	$w_{1,5}$	$w_{2,1}$	$w_{2,3}$	$w_{2,4}$	$w_{2,5}$	$w_{3,1}$	$w_{3,2}$	$w_{3,4}$	$w_{3,5}$	$w_{4,1}$	$w_{4,3}$	$w_{4,5}$
100	5.014	2.508	0.054	2.612	0.001	1.221	0.529	1.144	0.000	0.361	0.276	6.137	0.001	0.001	0.438
110	5.490	1.931	0.038	2.414	0.003	0.969	0.435	0.807	0.000	0.450	0.238	4.581	0.001	0.002	0.496
120	5.211	1.390	0.028	1.970	0.023	0.754	0.359	0.568	0.000	0.466	0.203	3.360	0.001	0.003	0.511

Moreover, by selecting $\delta_\tau = 0.1$, and $\delta_{cv} = 0.2$, we discover that the two largest increases in WTP are obtained when $w_{1,2}$ or $w_{3,5}$ is improved (see Table 4.20). Therefore, the waiting times between Steps 1 and 2 and between Steps 3 and 5 are the BN-wtp as well. Reducing the mean waiting time and variability of these two steps could lead to improvement in all performance measures: mean and coefficient of variation of total waiting time and waiting-time performance.

4.6 Conclusions

This chapter introduces a system-theoretic method to study the diagnosis-to-surgery process for lung cancer patients. It decomposes the complex process into multiple serial ones, and derives closed formulas to evaluate the mean, the coefficient of variation, and the distribution of waiting time. Moreover, indicators to identify the mean waiting time and variability bottlenecks are developed as well. Such a method provides lung cancer specialists and caregivers a quantitative tool to study and improve the lung cancer diagnosis-to-treatment process, and can also be applied to similar processes for other diseases.

Chapter 5

Reducing Delays in Inpatient Transitions

5.1 Introduction

Understanding the complete patient flow throughout the hospital is crucial for hospital administrators in predicting the expected demand and planning resources appropriately. Specifically, safe and efficient patient transitions are of critical importance to ensure patient safety and quality. In this chapter, a quantitative model to study the patient transitions within the hospital by taking into account the interactions among different units is presented. The system under study is described in Section 5.2, and is formulated as a finite capacity queueing network model in Section 5.3. As the arrival and departure processes are unknown, an iterative procedure for performance evaluation is introduced in Section 5.4. The convergence of the procedure and accuracy of estimation, as well as computation efficiency, are addressed in Section 5.5, and the impact of unit capacity, admission, and variability changes are discussed in Section 5.6. Finally, conclusions are formulated in Section 5.7.

5.2 System Description

To describe the system, the following notations are used and summarized in Table 5.21.

Table 5.21: Notations

(a) Service process

$c_{\text{ed}}, c_{\text{icu}}, c_{\text{wd}}$	number of beds in ED, ICU and ward
$T_{\text{ed}}, T_{\text{icu}}, T_{\text{wd}}$	service process in ED, ICU and ward
$\mu_{\text{ed}}, \mu_{\text{icu}}, \mu_{\text{wd}}$	service rate in ED, ICU and ward
$\tau_{\text{ed}}, \tau_{\text{icu}}, \tau_{\text{wd}}$	mean service time in ED, ICU and ward
$CV_{\text{ed}}, CV_{\text{icu}}, CV_{\text{wd}}$	CV of service time in ED, ICU and ward

(b) Arrival process

$T_{\text{arr}_{\text{ed}}}, T_{\text{arr}_{\text{icu}}}, T_{\text{arr}_{\text{wd}}}$	arrival process to ED, ICU and ward
$\lambda_{\text{ed}}, \lambda_{\text{icu}}, \lambda_{\text{wd}}$	arrival rate to ED, ICU and ward
$CV_{\text{arr}_{\text{ed}}}, CV_{\text{arr}_{\text{icu}}}, CV_{\text{arr}_{\text{wd}}}$	CV of arrival process to ED, ICU, ward
$\lambda_{\text{ext,ed}}, \lambda_{\text{ext,wd}}$	external arrival rate to ED and ward
$\lambda_{\text{ed,icu}}, \lambda_{\text{wd,icu}}$	arrival rate to ICU from ED and ward
$\lambda_{\text{ed,wd}}, \lambda_{\text{icu,wd}}$	arrival rate to ward from ED and ICU

(c) Departure process

$T_{\text{dc}_{\text{ed}}}, T_{\text{dc}_{\text{icu}}}, T_{\text{dc}_{\text{wd}}}$	departure process in ED, ICU, ward
$CV_{\text{dc}_{\text{ed}}}, CV_{\text{dc}_{\text{icu}}}, CV_{\text{dc}_{\text{wd}}}$	CV of departure process in ED, ICU and ward

(d) Transitions	
$p_{\text{ed,icu}}, p_{\text{ed,wd}}$	transition probability from ED to ICU and ward
$p_{\text{wd,icu}}, p_{\text{icu,wd}}$	transition probability between ward and ICU
$p_{\text{ed,hm}}, p_{\text{wd,hm}}$	discharge probability to home from ED and ward
(e) Waiting process	
$q_{\text{ed,wd}}, q_{\text{ed,icu}}$	average waiting time in ED for transfers to ward and ICU in M/M/c queue
$q_{\text{wd,icu}}, q_{\text{icu,wd}}$	average waiting time in ward for transfers to ICU and vice versa in M/M/c queue
$Q_{\text{ed,wd}}, Q_{\text{ed,icu}}$	waiting process in ED for transfers to ward, ICU in M/M/c queue
$Q_{\text{wd,icu}}, Q_{\text{icu,wd}}$	waiting process in ward for transfers to ICU and vice versa in M/M/c queue
$\widetilde{q}_{\text{ed,wd}}, \widetilde{q}_{\text{ed,icu}}$	average waiting time in ED for transfers to ward and ICU in G/G/c queue
$\widetilde{q}_{\text{wd,icu}}, \widetilde{q}_{\text{icu,wd}}$	average waiting time in ward for transfers to ICU and vice versa in G/G/c queue
(f) Departments	
$\rho_{\text{ed}}, \rho_{\text{icu}}, \rho_{\text{wd}}$	utilization of ED, ICU and ward
$\rho_{\text{icu}_{\text{ed}}}, \rho_{\text{icu}_{\text{wd}}}$	utilization of ICU for patients from ED and ward
$\rho_{\text{wd}_{\text{ed}}}, \rho_{\text{wd}_{\text{icu}}}$	utilization of ward for patients from ED and ICU

π_{ed},π_{icu},π_{wd}	probability of full in ED, ICU and ward
$\widetilde{\tau}_{\text{ed}}, \widetilde{\tau}_{\text{icu}}, \widetilde{\tau}_{\text{wd}}$	effective service time (including waiting time) or length of stay at ED, ICU, ward
$\widetilde{Var}_{\text{ed}}, \widetilde{Var}_{\text{icu}}, \widetilde{Var}_{\text{wd}}$	variance of effective service time at ED, ICU and ward
$\widetilde{CV}_{\text{ed}}, \widetilde{CV}_{\text{icu}}, \widetilde{CV}_{\text{wd}}$	CV of effective service time at ED, ICU and ward

The patient transitions among ED, ICU, and general ward within a hospital are illustrated in Figure 5.14. The following assumptions describe these transitions and the associated parameters.

i) Three major units in a hospital are considered: emergency department, intensive care unit, and general ward. There are c_{ed}, c_{icu}, and c_{wd} number of beds available in ED, ICU, and ward, respectively.

ii) The external arrival processes for ED and ward admissions, T_{arr_i}, i = ed, wd, follow general distribution with average number of arrivals per unit of time $\lambda_{\text{ext,ed}}$ and $\lambda_{\text{ext,wd}}$, respectively. In addition, the coefficients of variation of inter-arrival times are $CV_{\text{ext,ed}}$ and $CV_{\text{ext,wd}}$, respectively. Here subscript "ext" indicates that arrivals are external.

iii) The service processes, T_{i}, i = ed, wd, icu, are independent of priorities and arrivals. The service times follow general distributions with mean τ_{ed}, τ_{wd}, τ_{icu}, and coefficients of variation CV_{ed}, CV_{wd}, CV_{icu}, in ED, ward, and ICU, respectively.

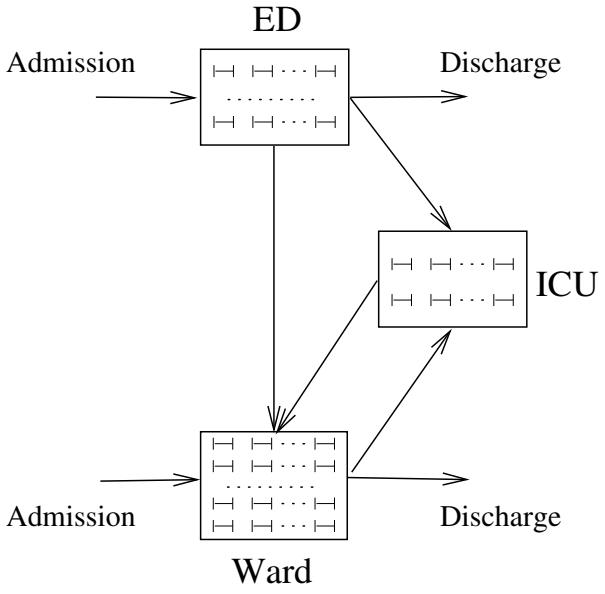

Figure 5.14: Patient transitions within hospital

iv) The routing probabilities between the units are defined as:

$$\begin{bmatrix} p_{\text{ed,dc}} & - & p_{\text{ed,wd}} & p_{\text{ed,icu}} \\ p_{\text{wd,dc}} & - & - & p_{\text{wd,icu}} \\ - & - & p_{\text{icu,wd}} & - \end{bmatrix}$$

v) For transitions into ward, priority is given in the order of patients transferring from ICU, ED, and new external admissions. For transitions into ICU, priority is given in the order of patients transiting from ward than ED.

vi) Treatment during transfers, mortality and diversion to other hospitals are not considered.

Remark 5.1 For simplification purpose, subscripts "ed", "icu", "wd", "arr" and "dc" are used to denote ED, ICU, ward, admission, and discharge, respectively. In addition, "arr_{icu}", "arr_{ed}" and "arr_{wd}", represent the overall arrival processes to ICU, ED,

and ward, respectively, and "dc_{icu}", "dc_{ed}", "dc_{wd}" characterize the departure processes from ICU, ED, and ward, respectively. □

Remark 5.2 The reason why external admissions are given the lowest priority is that normally, as bed occupancy level rises, priority is given to admitting the urgent patients from the internal units while pre-scheduled admissions need to be canceled. Among the two internal units ED and ICU, a higher priority is given to ICU since keeping a patient in an ICU bed is more costly than placing in a ward bed, especially when this patient no longer needs intensive care. □

Remark 5.3 Although the exact performance is dependent on the complete distribution of service time, when the variability is not large, the system performance primarily relies on the mean and coefficient of variation. Such a property has been verified in many healthcare and manufacturing studies (e.g., [53, 94, 51, 98] and [93]), and is also implied in G/G/c model. □

5.3 Queueing Network Model

Considering each unit (ED, ICU or ward), a queueing model can be established to characterize the arrival, service and departure processes as shown in Figure 5.15 where the admission queues are outside of the units and the departure queues are within the units.

If all units have unlimited capacity, a patient will be immediately served and the patient's length of stay in a unit will be the same as its service time. That is, we

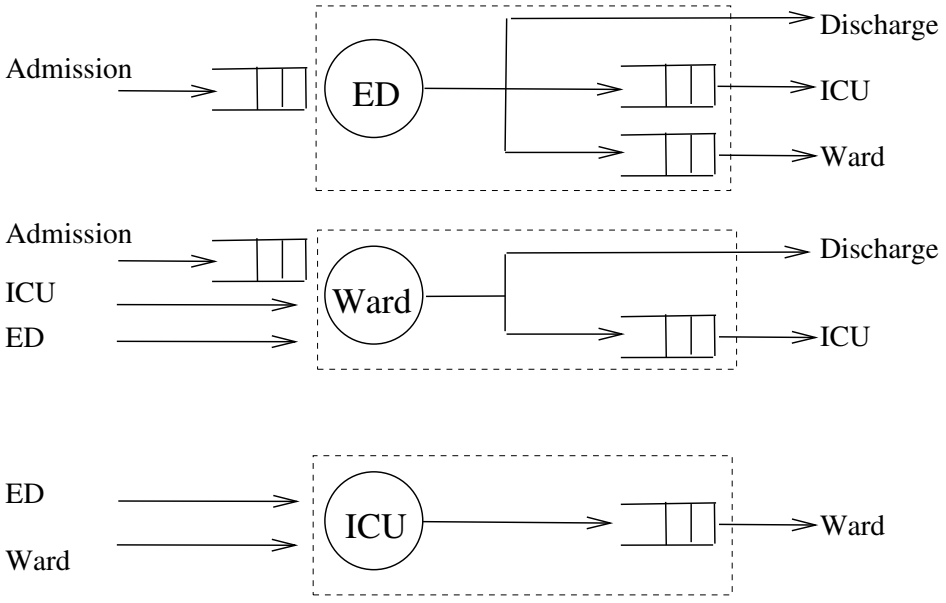

Figure 5.15: Queueing model of patient transitions

can decouple the service processes and analyze each unit independently. For instance, the ED can be analyzed without considering boarding to other units. However, in real life scenarios, bed capacity is limited and blockings are commonly observed. Thus, additional waiting time will be added to a patient's length of stay in a unit. This will affect the departure process of the unit and the arrival process of the destination unit. Thus, blocking at one unit affects all other units, which makes it impossible to analyze each unit independently. To solve this problem, an iteration method is proposed.

The idea of the iteration procedure is illustrated in Figure 5.16 and outlined below.

For each unit, we assume general arrival with rate λ_i, $i = \mathrm{ed}, \mathrm{icu}, \mathrm{wd}$. When no blocking occurs, the departure process can be analyzed as follows:

Proposition 5.1 *Under assumptions (i)-(vi), when there is no blocking, the arrivals*

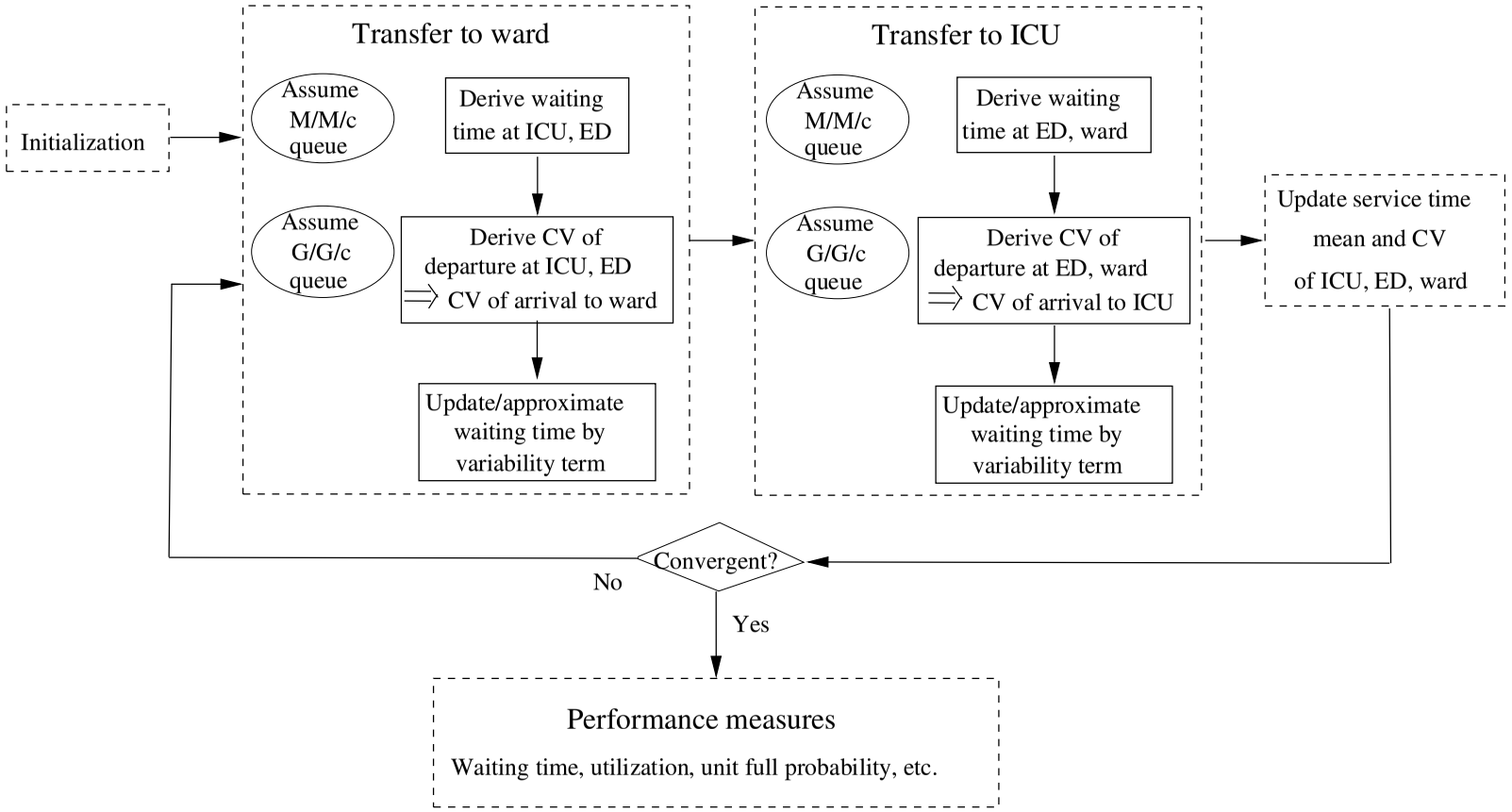

Figure 5.16: Illustration of iteration procedure

for each unit can be evaluated as:

$$\begin{aligned}
\lambda_{\text{ed}} &= \lambda_{\text{ext,ed}}, \\
\lambda_{\text{icu}} &= \frac{\lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,icu}}}{1 - p_{\text{icu,wd}} p_{\text{wd,icu}}}, \\
\lambda_{\text{wd}} &= \frac{\lambda_{\text{ext,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}}}{1 - p_{\text{icu,wd}} p_{\text{wd,icu}}}.
\end{aligned} \tag{5.1}$$

In addition, the arrivals from specific units are:

$$\lambda_{\text{ed,icu}} = \lambda_{\text{ext,ed}} p_{\text{ed,icu}},$$

$$\lambda_{\text{ed,wd}} = \lambda_{\text{ext,ed}} p_{\text{ed,wd}},$$

$$\begin{aligned}
\lambda_{\text{wd,icu}} = [&\lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&+ \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}} p_{\text{wd,icu}}]/[1 - p_{\text{icu,wd}} p_{\text{wd,icu}}],
\end{aligned} \tag{5.2}$$

$$\begin{aligned}
\lambda_{\text{icu,wd}} = [&\lambda_{\text{ext,wd}} p_{\text{wd,icu}} p_{\text{icu,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}} \\
&+ \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} p_{\text{icu,wd}}]/[1 - p_{\text{icu,wd}} p_{\text{wd,icu}}].
\end{aligned} \tag{5.3}$$

Proof: See Appendix C. ■

As there are multiple arrival sources with different priorities for ward and ICU, using a non-preemptive multi-server system with multiple priority classes would be an option. However, as calculating the variability of waiting times for $G/G/c$ queues is extremely difficult, we introduce an approximation approach. First, we calculate the first two moments of the waiting times by assuming an $M/M/c$ queuing model. Then we adjust the waiting times to take into account the variabilities to approximate the $G/G/c$ queue.

First, select either the ward or ICU to start analyzing the patient flow. Note that since we are only interested in internal blockings (i.e., not external admission waiting times which occur outside the hospital), patient admissions to ED are not considered.

Thus, considering ward first, we assume an $M/M/c_{wd}$ queue for ward with capacity c_{wd}. If we know the mean inter-arrival time and the mean service time, we can derive the probability the unit is full (π_{wd}) and the first two moments of waiting times for patients coming from ED and ICU ($q_{i,\text{wd}}$ and $E[(Q_{i,\text{wd}})^2]$, $i =$ ed, icu). Specifically, we introduce a probabilistic equivalence between the $M/G/c$ queue with multiple servers' vacations and the $M/M/c$ system to obtain the first two moments of queue times [99].

Next, to approximate the waiting times for $G/G/c$ queue, Kingman's approximation formula is used. The waiting time at $G/G/c$ queue can be obtained by multiplying a variability term to the waiting time obtained from $M/M/c$ queue. Here, the variability term consists of the CVs of both the arrival process and the service process. Note that for the internal arrivals, departures from one unit form the arrivals to other units, that is, the CV of departure process from unit A to destination unit B is equivalent to the CV of arrival process flowing into unit B from unit A. Particularly, as patients from both ED and ICU can transfer to ward, the CVs of the departure processes from both ED and ICU, $CV_{\text{ed,wd}}$ and $CV_{\text{icu,wd}}$, need to be known, which can be calculated as follows:

Assuming infinite buffer, the departure variability depends on the arrival variability ($CV_{\text{arr},i}$) and process variability (CV_i) where the two terms are weighted by the square of the utilization of the server (ρ_i). Following this approach, the variability of the discharge process at ED and ICU can be derived, which is equivalent to the variability of the arrival process to ward from ED and ICU, i.e.,

$$CV^2_{\text{dc}_i} = \rho_i^2 CV_i^2 + (1 - \rho_i^2) CV^2_{\text{arr},i}, \quad i = \text{ed}, \text{icu}.$$

By linking these two arrival processes with the external admission to ward, we obtain

the total arrival process to ward, i.e.,

$$CV^2_{\text{arr}_{\text{wd}}} = \frac{\frac{CV^2_{\text{icu,wd}}}{\lambda^2_{\text{icu,wd}}} + \frac{CV^2_{\text{ed,wd}}}{\lambda^2_{\text{ed,wd}}} + \frac{CV^2_{\text{ext,wd}}}{\lambda^2_{\text{ext,wd}}}}{\left(\frac{1}{\lambda_{\text{icu,wd}}} + \frac{1}{\lambda_{\text{ed,wd}}} + \frac{1}{\lambda_{\text{ext,wd}}}\right)^2}.$$

Now, by multiplying the waiting times obtained from $M/M/c$ queue with the variability term, we can approximate the real waiting times in $G/G/c$ queue.

$$\widetilde{q}_{i,\text{wd}} = \frac{CV^2_{i,\text{wd}} + CV^2_{\text{wd}}}{2} \cdot q_{i,\text{wd}}, \quad i = \text{ed}, \text{icu}.$$

Analogously, similar analyses can be applied to ICU. Again, first an $M/M/c_{icu}$ queue is assumed, and waiting process $q_{i,\text{icu}}$ and $E[(Q_{i,\text{icu}})^2]$, $i = \text{wd}$, ed, can be derived. Using departure processes at ED and ward, the arrivals to ICU, $CV_{\text{arr}_{\text{icu}}}$, can be obtained. Then the waiting times to ICU for patients transiting from ED and ward in $G/G/c_{\text{icu}}$ queue, $\widetilde{q}_{i,\text{icu}}$, $i = \text{ed}$, wd, can be evaluated.

Finally, by adding the waiting times at ED, ICU, and ward to their corresponding service times, we obtain the effective service process in each unit (i.e., the unit's length of stay),

$$\widetilde{\tau}_i = \tau_i + p_{i,j}\widetilde{q}_{i,j}, \quad i, j = \text{wd}, \text{icu}, i \neq j,$$

$$\widetilde{\tau}_{\text{ed}} = \tau_{\text{ed}} + \sum_{j=\text{wd,icu}} p_{\text{ed},j}\widetilde{q}_{\text{ed},j}.$$

In addition to the average time, the variability can also be calculated.

$$\widetilde{Var}_i = Var_i + p_{i,j}\widetilde{Var}_{i,j}, \quad i, j = \text{wd}, \text{icu}, i \neq j,$$

$$\widetilde{Var}_{\text{ed}} = Var_{\text{ed}} + \sum_{j=\text{wd,icu}} p_{\text{ed},j}\widetilde{Var}_{\text{ed},j}.$$

However, since both the arrival and departure processes are unknown, we introduce an iterative approach. First, assuming internal arrivals follow Poisson distribution, we

derive the waiting time mean and CV and accommodate them into the service time. Replacing the mean and CV of each service process, we go back to re-evaluate the waiting time and repeat the procedure. When the procedure converges, we obtain the approximation of the mean and CV of waiting times.

5.4 Iteration Procedure

Formally, the iteration procedure can be described as follows:

5.4.1 Step 1: Initialization

Assume

$$\begin{aligned}
\widetilde{\tau}_i^{(0)} &= \tau_i, \quad i = \text{ed}, \text{wd}, \text{icu}, \\
\widetilde{CV}_i^{(0)} &= CV_i, \quad i = \text{ed}, \text{wd}, \text{icu}, \\
CV_{\text{arr}_{\text{icu}}}^{(0)} &= 1, \\
\widetilde{q}_{i,\text{wd}}^{(0)} &= 0, \quad i = \text{ed}, \text{icu}, \\
\widetilde{q}_{i,\text{icu}}^{(0)} &= 0, \quad i = \text{ed}, \text{wd}.
\end{aligned}$$

Set $k = 0$.

5.4.2 Step 2: Update Transition Rates

Step 2.1: Waiting time to ward in $M/M/c_{\text{wd}}$ queues

With ward capacity c_{wd}, the first two moments of waiting times are derived by assuming $M/M/c_{\text{wd}}$ queues first.

From ICU to ward: For class 1 patients, i.e., patients from ICU to ward, the service rate and utilization can be calculated as

$$\begin{aligned}
\mu_{\text{wd}}^{(k+1)} &= \frac{1}{\widetilde{\tau}_{\text{wd}}^{(k)}}, \\
\rho_{\text{wd}}^{(k+1)} &= \frac{\lambda_{\text{wd}}}{c_{\text{wd}} \cdot \mu_{\text{wd}}^{(k+1)}}, \\
\rho_{\text{wd}_{\text{icu}}}^{(k+1)} &= \frac{p_{\text{icu,wd}} \cdot \lambda_{\text{icu}}}{c_{\text{wd}} \cdot \mu_{\text{wd}}^{(k+1)}}.
\end{aligned} \tag{5.4}$$

In addition, define parameters γ and π as

$$\begin{aligned}
\gamma_{\text{wd}_{\text{icu}}}^{(k+1)} &= \rho_{\text{wd}_{\text{icu}}}^{(k+1)}, \\
\pi_{\text{wd}}^{(k+1)} &= \frac{\dfrac{(\lambda_{\text{wd}}/\mu_{\text{wd}}^{(k+1)})^{c_{\text{wd}}}}{c_{\text{wd}}!(1-\rho_{\text{wd}}^{(k+1)})}}{\sum_{i=0}^{c_{\text{wd}}-1} \dfrac{(\lambda_{\text{wd}}/\mu_{\text{wd}}^{(k+1)})^{i}}{i!} + \dfrac{(\lambda_{\text{wd}}/\mu_{\text{wd}}^{(k+1)})^{c_{\text{wd}}}}{c_{\text{wd}}!(1-\rho_{\text{wd}}^{(k+1)})}},
\end{aligned} \tag{5.5}$$

where π_i represents the probability unit i is full. Then the first and second moments of waiting times can be derived as

$$\begin{aligned}
q_{\text{icu,wd}}^{(k+1)} &= \frac{\pi_{\text{wd}}^{(k+1)}}{c_{\text{wd}}\mu_{\text{wd}}^{(k+1)}(1-\gamma_{\text{wd}_{\text{icu}}}^{(k+1)})}, \\
E[(Q_{\text{icu,wd}}^{(k+1)})^2] &= \frac{2\pi_{\text{wd}}^{(k+1)}}{(c_{\text{wd}}\mu_{\text{wd}}^{(k+1)})^2(1-\gamma_{\text{wd}_{\text{icu}}}^{(k+1)})^2}.
\end{aligned} \tag{5.6}$$

From ED to ward: For class 2 patients coming from ED to ward, define

$$\begin{aligned}
\rho_{\text{wd}_{\text{ed}}}^{(k+1)} &= \frac{p_{\text{ed,wd}} \cdot \lambda_{\text{ed}}}{c_{\text{wd}} \cdot \mu_{\text{wd}}^{(k+1)}}, \\
\gamma_{\text{wd}_{\text{ed}}}^{(k+1)} &= \rho_{\text{wd}_{\text{icu}}}^{(k+1)} + \rho_{\text{wd}_{\text{ed}}}^{(k+1)}.
\end{aligned} \tag{5.7}$$

Then we obtain

$$q_{\mathrm{ed,wd}}^{(k+1)} = \frac{\pi_{\mathrm{wd}}^{(k+1)}}{c_{\mathrm{wd}}\mu_{\mathrm{wd}}^{(k+1)}(1-\gamma_{\mathrm{wd_{ed}}}^{(k+1)})(1-\gamma_{\mathrm{wd_{icu}}}^{(k+1)})}, \tag{5.8}$$
$$E[(Q_{\mathrm{ed,wd}}^{(k+1)})^2] = \frac{2\pi_{\mathrm{wd}}^{(k+1)}(1-\gamma_{\mathrm{wd_{ed}}}^{(k+1)}\gamma_{\mathrm{wd_{icu}}}^{(k+1)})}{(c_{\mathrm{wd}}\mu_{\mathrm{wd}}^{(k+1)})^2(1-\gamma_{\mathrm{wd_{ed}}}^{(k+1)})^2(1-\gamma_{\mathrm{wd_{icu}}}^{(k+1)})^3}.$$

Step 2.2: Waiting time to ward in $G/G/c_{\mathrm{wd}}$ queues

For arrivals from ED, the service rate and utilization at ED are represented by

$$\mu_{\mathrm{ed}}^{(k+1)} = \frac{1}{\widetilde{\tau}_{\mathrm{ed}}^{(k)}},$$
$$\rho_{\mathrm{ed}}^{(k+1)} = \frac{\lambda_{\mathrm{ed}}}{c_{\mathrm{ed}}\mu_{\mathrm{ed}}^{(k+1)}}. \tag{5.9}$$

Then the discharge process variability at ED is described by

$$(CV_{\mathrm{dc_{ed}}}^{(k+1)})^2 = (\rho_{\mathrm{ed}}^{(k+1)})^2(\widetilde{CV}_{\mathrm{ed}}^{(k)})^2 + [1-(\rho_{\mathrm{ed}}^{(k+1)})^2](CV_{\mathrm{ext,ed}})^2. \tag{5.10}$$

Similarly, for arrivals from ICU, we have

$$\mu_{\mathrm{icu}}^{(k+1)} = \frac{1}{\widetilde{\tau}_{\mathrm{icu}}^{(k)}},$$
$$\rho_{\mathrm{icu}}^{(k+1)} = \frac{\lambda_{\mathrm{icu}}}{c_{\mathrm{icu}}\mu_{\mathrm{icu}}^{(k+1)}}. \tag{5.11}$$

Then the discharge variability at ICU is evaluated by

$$(CV_{\mathrm{dc_{icu}}}^{(k+1)})^2 = (\rho_{\mathrm{icu}}^{(k+1)})^2(\widetilde{CV}_{\mathrm{icu}}^{(k)})^2 + [1-(\rho_{\mathrm{icu}}^{(k+1)})^2](CV_{\mathrm{arr_{icu}}}^{(k)})^2. \tag{5.12}$$

By linking the departure process from the originating unit to the arrival process to the destination unit, the arrivals from ICU and ED are evaluated as:

$$(CV^{(k+1)}_{\text{icu,wd}})^2 = (CV^{(k+1)}_{\text{dc}_{\text{icu}}})^2,$$
$$(CV^{(k+1)}_{\text{ed,wd}})^2 = (CV^{(k+1)}_{\text{dc}_{\text{ed}}})^2.$$

Using $\lambda_{\text{icu,wd}}$ and $\lambda_{\text{ed,wd}}$ from Proposition 5.1, the variability of the total arrival process to ward can be derived as:

$$(CV^{(k+1)}_{\text{arr}_{\text{wd}}})^2 = \frac{\left(\frac{CV^{(k+1)}_{\text{icu,wd}}}{\lambda_{\text{icu,wd}}}\right)^2 + \left(\frac{CV^{(k+1)}_{\text{ed,wd}}}{\lambda_{\text{ed,wd}}}\right)^2 + \left(\frac{CV_{\text{ext,wd}}}{\lambda_{\text{ext,wd}}}\right)^2}{\left(\frac{1}{\lambda_{\text{icu,wd}}} + \frac{1}{\lambda_{\text{ed,wd}}} + \frac{1}{\lambda_{\text{ext,wd}}}\right)^2}. \tag{5.13}$$

Then, the waiting times for $G/G/c$ queue, $\widetilde{q}$ can be approximated by multiplying the wait time in an analogous $M/M/c$ queue, q, by the variability terms:

$$\widetilde{q}^{(k+1)}_{\text{icu,wd}} = \frac{(CV^{(k+1)}_{\text{icu,wd}})^2 + (\widetilde{CV}^{(k)}_{\text{wd}})^2}{2} \cdot q^{(k+1)}_{\text{icu,wd}},$$
$$\widetilde{q}^{(k+1)}_{\text{ed,wd}} = \frac{(CV^{(k+1)}_{\text{ed,wd}})^2 + (\widetilde{CV}^{(k)}_{\text{wd}})^2}{2} \cdot q^{(k+1)}_{\text{ed,wd}}. \tag{5.14}$$

Step 2.3: Waiting time to ICU in $M/M/c_{\text{icu}}$ queues

For ICU, first assume $M/M/c_{\text{icu}}$ queue to derive the first two moments of waiting times.

From ward to ICU: For ward to ICU patients (class 1), we have

$$\rho^{(k+1)}_{\text{icu}_{\text{wd}}} = \frac{p_{\text{wd,icu}} \cdot \lambda_{\text{wd}}}{c_{\text{icu}} \cdot \mu^{(k+1)}_{\text{icu}}},$$
$$\gamma^{(k+1)}_{\text{icu}_{\text{wd}}} = \rho^{(k+1)}_{\text{icu}_{\text{wd}}}, \tag{5.15}$$
$$\pi^{(k+1)}_{\text{icu}} = \frac{\frac{(\lambda_{\text{icu}}/\mu^{(k+1)}_{\text{icu}})^{c_{\text{icu}}}}{c_{\text{icu}}!(1-\rho^{(k+1)}_{\text{icu}})}}{\sum_{i=0}^{c_{\text{icu}}-1} \frac{(\lambda_{\text{icu}}/\mu^{(k+1)}_{\text{icu}})^i}{i!} + \frac{(\lambda_{\text{icu}}/\mu^{(k+1)}_{\text{icu}})^{c_{\text{icu}}}}{c_{\text{icu}}!(1-\rho^{(k+1)}_{\text{icu}})}}.$$

Thus, the waiting times are characterized by

$$q_{\text{wd,icu}}^{(k+1)} = \frac{\pi_{\text{icu}}^{(k+1)}}{c_{\text{icu}}\mu_{\text{icu}}^{(k+1)}(1-\gamma_{\text{icu}_{\text{wd}}}^{(k+1)})}, \tag{5.16}$$
$$E[(Q_{\text{wd,icu}}^{(k+1)})^2] = \frac{2\pi_{\text{icu}}^{(k+1)}}{(c_{\text{icu}}\mu_{\text{icu}}^{(k+1)})^2(1-\gamma_{\text{icu}_{\text{wd}}}^{(k+1)})^2}.$$

From ED to ICU: For ED to ICU patients (class 2), from

$$\rho_{\text{icu}_{\text{ed}}}^{(k+1)} = \frac{p_{\text{ed,icu}} \cdot \lambda_{\text{ed}}}{c_{\text{icu}} \cdot \mu_{\text{icu}}^{(k+1)}},$$
$$\gamma_{\text{icu}_{\text{ed}}}^{(k+1)} = \rho_{\text{icu}_{\text{wd}}}^{(k+1)} + \rho_{\text{icu}_{\text{ed}}}^{(k+1)}. \tag{5.17}$$

we obtain

$$q_{\text{ed,icu}}^{(k+1)} = \frac{\pi_{\text{icu}}^{(k+1)}}{c_{\text{icu}}\mu_{\text{icu}}^{(k+1)}(1-\gamma_{\text{icu}_{\text{ed}}}^{(k+1)})(1-\gamma_{\text{icu}_{\text{wd}}}^{(k+1)})}, \tag{5.18}$$
$$E[(Q_{\text{ed,icu}}^{(k+1)})^2] = \frac{2\pi_{\text{icu}}^{(k)}(1-\gamma_{\text{icu}_{\text{ed}}}^{(k+1)}\gamma_{\text{icu}_{\text{wd}}}^{(k+1)})}{(c_{\text{icu}}\mu_{\text{icu}}^{(k+1)})^2(1-\gamma_{\text{icu}_{\text{ed}}}^{(k+1)})^2(1-\gamma_{\text{icu}_{\text{wd}}}^{(k+1)})^3}.$$

Step 2.4: Waiting time to ICU in $G/G/c_{\text{icu}}$ queues

As patients from both ward and ED transfer to ICU, the departure processes from both ward and ED need to be known. To measure flow variability, assume infinite buffers and calculate departure variability.

$$(CV_{\text{dc}_{\text{wd}}}^{(k+1)})^2 = (\rho_{\text{wd}}^{(k+1)})^2(\widetilde{CV}_{\text{wd}}^{(k)})^2 + (1-(\rho_{\text{wd}}^{(k+1)})^2)(CV_{\text{arr}_{\text{wd}}}^{(k+1)})^2. \tag{5.19}$$

Using the departure variability, we can calculate the CV of inter-arrival time for each patient class.

$$(CV_{\text{wd,icu}}^{(k+1)})^2 = (CV_{\text{dc}_{\text{wd}}}^{(k+1)})^2,$$
$$(CV_{\text{ed,icu}}^{(k+1)})^2 = (CV_{\text{dc}_{\text{ed}}}^{(k+1)})^2.$$

Using $\lambda_{\text{wd,icu}}$ and $\lambda_{\text{ed,icu}}$ from Proposition 5.1, it implies that

$$(CV_{\text{arr}_{\text{icu}}}^{(k+1)})^2 = \frac{\left(\frac{CV_{\text{wd,icu}}^{(k+1)}}{\lambda_{\text{wd,icu}}}\right)^2 + \left(\frac{CV_{\text{ed,icu}}^{(k+1)}}{\lambda_{\text{ed,icu}}}\right)^2}{\left(\frac{1}{\lambda_{\text{wd,icu}}} + \frac{1}{\lambda_{\text{ed,icu}}}\right)^2}. \tag{5.20}$$

Then, we approximate the waiting times for $G/G/c_{\text{icu}}$ queue by multiplying the waiting time in the analogous $M/M/c_{\text{icu}}$ queue by the variability term.

$$\begin{aligned}
\widetilde{q}_{\text{wd,icu}}^{(k+1)} &= \frac{(CV_{\text{wd,icu}}^{(k+1)})^2 + (\widetilde{CV}_{\text{icu}}^{(k)})^2}{2} \cdot q_{\text{wd,icu}}^{(k+1)}, \\
\widetilde{q}_{\text{ed,icu}}^{(k+1)} &= \frac{(CV_{\text{ed,icu}}^{(k+1)})^2 + (\widetilde{CV}_{\text{icu}}^{(k)})^2}{2} \cdot q_{\text{ed,icu}}^{(k+1)}.
\end{aligned} \tag{5.21}$$

Step 2.5: Service time at ward

Using the above results, we now update the first two moments of effective service times at ward, where the service time and queueing time are assumed to be independent.

$$\begin{aligned}
\widetilde{\tau}_{\text{wd}}^{(k+1)} &= \tau_{\text{wd}} + p_{\text{wd,icu}} \cdot \widetilde{q}_{\text{wd,icu}}^{(k+1)}, \\
\widetilde{Var}_{\text{wd}}^{(k+1)} &= (CV_{\text{wd}}\tau_{\text{wd}})^2 + p_{\text{wd,icu}}^2 \cdot (E[(Q_{\text{wd,icu}}^{(k+1)})^2] \\
&\quad - (q_{\text{wd,icu}}^{(k+1)})^2), \\
\widetilde{CV}_{\text{wd}}^{(k+1)} &= \frac{\sqrt{\widetilde{Var}_{\text{wd}}^{(k+1)}}}{\widetilde{\tau}_{\text{wd}}^{(k+1)}}.
\end{aligned} \tag{5.22}$$

Step 2.6: Service time at ICU

Using the derived waiting times, we update the mean and standard deviation of effective service times at ICU.

$$\begin{aligned}
\widetilde{\tau}_{\text{icu}}^{(k+1)} &= \tau_{\text{icu}} + p_{\text{icu,wd}} \cdot \widetilde{q}_{\text{icu,wd}}^{(k+1)}, \\
\widetilde{Var}_{\text{icu}}^{(k+1)} &= (CV_{\text{icu}}\tau_{\text{icu}})^2 + p_{\text{icu,wd}}^2 \cdot (E[(Q_{\text{icu,wd}}^{(k+1)})^2] \\
&\quad - (q_{\text{icu,wd}}^{(k+1)})^2), \\
\widetilde{CV}_{\text{icu}}^{(k+1)} &= \frac{\sqrt{\widetilde{Var}_{\text{icu}}^{(k+1)}}}{\widetilde{\tau}_{\text{icu}}^{(k+1)}}.
\end{aligned} \tag{5.23}$$

Step 2.7: Service time at ED

Analogously, the mean and standard deviation of effective service times at ED are calculated.

$$\begin{aligned}
\widetilde{\tau}_{\text{ed}}^{(k+1)} &= \tau_{\text{ed}} + p_{\text{ed,wd}} \cdot \widetilde{q}_{\text{ed,wd}}^{(k+1)} + p_{\text{ed,icu}} \cdot \widetilde{q}_{\text{ed,icu}}^{(k+1)}, \\
\widetilde{Var}_{\text{ed}}^{(k+1)} &= (CV_{\text{ed}}\tau_{\text{ed}})^2 + p_{\text{ed,wd}}^2 \cdot (E[(Q_{\text{ed,wd}}^{(k+1)})^2] \\
&\quad - (q_{\text{ed,wd}}^{(k+1)})^2) + p_{\text{ed,icu}}^2 \cdot (E[(Q_{\text{ed,icu}}^{(k+1)})^2] \\
&\quad - (q_{\text{ed,icu}}^{(k+1)})^2), \\
\widetilde{CV}_{\text{ed}}^{(k+1)} &= \frac{\sqrt{\widetilde{Var}_{\text{ed}}^{(k+1)}}}{\widetilde{\tau}_{\text{ed}}^{(k+1)}}.
\end{aligned} \tag{5.24}$$

5.4.3 Step 3. Check Stopping Condition

Check iteration stopping condition, i.e., whether the maximal difference between two consecutive iterations in mean queue time is less than $\epsilon = 10^{-6}$ or not.

If $\max\{\epsilon_i, i = 1, \ldots, 4\} < \epsilon$, where

$$\epsilon_1 = |\widetilde{q}_{\text{icu,wd}}^{(k+1)} - \widetilde{q}_{\text{icu,wd}}^{(k)}|, \qquad \epsilon_2 = |\widetilde{q}_{\text{ed,wd}}^{(k+1)} - \widetilde{q}_{\text{ed,wd}}^{(k)}|,$$
$$\epsilon_3 = |\widetilde{q}_{\text{ed,icu}}^{(k+1)} - \widetilde{q}_{\text{ed,icu}}^{(k)}|, \qquad \epsilon_4 = |\widetilde{q}_{\text{wd,icu}}^{(k+1)} - \widetilde{q}_{\text{wd,icu}}^{(k)}|,$$

then iteration stops. Otherwise, set $k = k + 1$ and return to Step 2.

5.5 Performance Measures

5.5.1 Convergence

Through extensive numerical studies, we always observe that the iteration procedure described in Section 5.4 is convergent. An example is shown in Figure 5.17 to illustrate the convergence during the iteration process where the procedure converges in about 5 iterations. Among all the examples we tested, most of them converge within 5-10 iterations, and a few up to 15-20 iterations.

Remark 5.4 The rationale of the convergence of the procedure can be explained as follows: when the mean and coefficient of variation of service time increase, we should expect more blocking which leads to longer service time (as we add the blocking time to the service time) and more variability. The increased service time and variability in turn lead to an increase in unit full probability. Since such probabilities are bounded, it implies that convergence exists. At this point, a rigorous proof can only be derived for queueing time with respect to mean service time (i.e., q with respect to τ) in the Markovian scenario. However, extensive numerical experiments have validated that the procedure converges in all scenarios. □

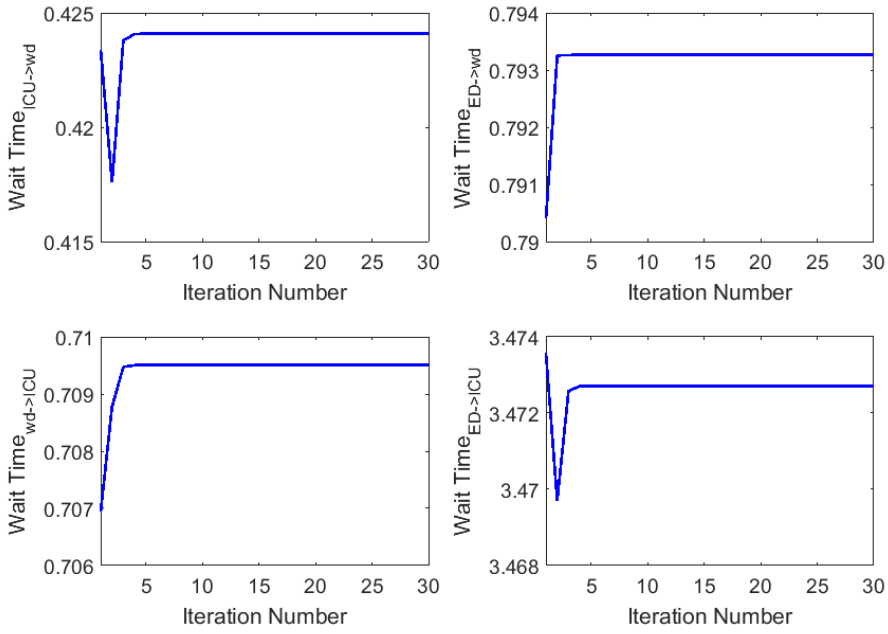

Figure 5.17: Illustration of convergence

Upon convergence, the system performance can be estimated.

$$
\begin{aligned}
\widetilde{q}_{i,\mathrm{wd}} &:= \lim_{k\to\infty} \widetilde{q}_{i,\mathrm{wd}}^{(k)}, \quad i = \mathrm{ed}, \mathrm{icu}, \\
\widetilde{q}_{i,\mathrm{icu}} &:= \lim_{k\to\infty} \widetilde{q}_{i,\mathrm{icu}}^{(k)}, \quad i = \mathrm{ed}, \mathrm{wd}, \\
\rho_i &:= \lim_{k\to\infty} \rho_i^{(k)}, \quad i = \mathrm{ed}, \mathrm{icu}, \mathrm{wd}, \\
\pi_i &:= \lim_{k\to\infty} \pi_i^{(k)}, \quad i = \mathrm{ed}, \mathrm{icu}, \mathrm{wd}, \\
\sigma_{i,\mathrm{icu}} &:= \lim_{k\to\infty} (E[(Q_{i,\mathrm{icu}}^{(k)})^2] - (q_{i,\mathrm{icu}}^{(k)})^2)^{\frac{1}{2}}, \quad i = \mathrm{ed}, \mathrm{wd}, \\
\sigma_{i,\mathrm{wd}} &:= \lim_{k\to\infty} (E[(Q_{i,\mathrm{wd}}^{(k)})^2] - (q_{i,\mathrm{wd}}^{(k)})^2)^{\frac{1}{2}}, \quad i = \mathrm{ed}, \mathrm{icu},
\end{aligned}
\tag{5.25}
$$

where $\sigma_{i,j}$ approximates the standard deviation of waiting times for transitions from unit i to unit j.

5.5.2 Accuracy

Upon convergence, the accuracy of the procedure is investigated numerically.

Parameter selection

Through extensive review of prevailing literatures, a reasonable range of system parameters are obtained.

- Bed capacity
  - Total beds: [285, 440]
  - Ward beds/ED beds: [4.25, 14.4]
  - ICU beds/ED beds: [1.5, 1.96]

    Note that actual ICU beds available to ED and ward patients are $55\% - 58\%$ of total ICU beds. The other half of bed capacity is occupied/reserved by operating theatre patients (either elective or emergency) and admissions from other hospitals.
  - Available ICU beds/ward beds: [0.82, 1.13]
- Transition probabilities
  - ED to ward: [10.3%, 21.2%]
  - ED to ICU: [1%, 5.2%]
  - Ward to ICU: [2%, 5%]
- Length of stay
  - ED: [3.68, 5.9] hours

  - Ward: [1.85, 6.98] days
  - ICU: [1.5, 5.5] days

- Admission rates
  - ED admission/ward admission: [2, 7.87]

- In addition, the variability measures, i.e., the CVs, of service and arrival processes are randomly selected between 0 and 1.

Experiments

Extensive numerical experiments have been carried out by randomly selecting parameters from the ranges and comparing with simulation results. In each simulation experiment, 10,000 time units are used for data collection period with 1,000 time units of warm-up period and 20 replications. Gamma and log-normal distributions are randomly selected for service and inter-arrival times in simulation experiments. The accuracy of the performance measures are defined as follows. For relative error,

$$\delta_{\widetilde{q}_{i,\mathrm{wd}}} = \frac{|\widetilde{q}^{proc}_{i,\mathrm{wd}} - \widetilde{q}^{sim}_{i,\mathrm{wd}}|}{\widetilde{q}^{sim}_{i,\mathrm{wd}}} \cdot 100\%, \quad i = \mathrm{ed}, \mathrm{icu},$$
$$\delta_{\widetilde{q}_{i,\mathrm{icu}}} = \frac{|\widetilde{q}^{proc}_{i,\mathrm{icu}} - \widetilde{q}^{sim}_{i,\mathrm{icu}}|}{\widetilde{q}^{sim}_{i,\mathrm{icu}}} \cdot 100\%, \quad i = \mathrm{ed}, \mathrm{wd},$$

where superscripts "sim" and "proc" indicate that the measures are obtained from simulation and iteration procedure, respectively. For absolute error

$$\Delta_{\widetilde{q}_{i,\mathrm{wd}}} = |\widetilde{q}^{proc}_{i,\mathrm{wd}} - \widetilde{q}^{sim}_{i,\mathrm{wd}}|, \quad i = \mathrm{ed}, \mathrm{icu},$$
$$\Delta_{\widetilde{q}_{i,\mathrm{icu}}} = |\widetilde{q}^{proc}_{i,\mathrm{icu}} - \widetilde{q}^{sim}_{i,\mathrm{icu}}|, \quad i = \mathrm{ed}, \mathrm{wd}.$$

In addition to mean time, the accuracy of variability in waiting times is also defined.

$$\delta_{\sigma_{i,\mathrm{wd}}} = \frac{|\sigma^{proc}_{i,\mathrm{wd}} - \sigma^{sim}_{i,\mathrm{wd}}|}{\sigma^{sim}_{i,\mathrm{wd}}} \cdot 100\%, \quad i = \mathrm{ed}, \mathrm{icu},$$

$$\delta_{\sigma_{i,\mathrm{icu}}} = \frac{|\sigma^{proc}_{i,\mathrm{icu}} - \sigma^{sim}_{i,\mathrm{icu}}|}{\sigma^{sim}_{i,\mathrm{icu}}} \cdot 100\%, \quad i = \mathrm{ed}, \mathrm{wd},$$

$$\Delta_{\sigma_{i,\mathrm{wd}}} = |\sigma^{proc}_{i,\mathrm{wd}} - \sigma^{sim}_{i,\mathrm{wd}}|, \quad i = \mathrm{ed}, \mathrm{icu},$$

$$\Delta_{\sigma_{i,\mathrm{icu}}} = |\sigma^{proc}_{i,\mathrm{icu}} - \sigma^{sim}_{i,\mathrm{icu}}|, \quad i = \mathrm{ed}, \mathrm{wd},$$

The results are summarized in Table 5.22 for both mean and standard deviation of waiting times. As one can see, for average waiting times from ED to ward and ward to ICU, the relative errors are very small, within about 1% and 6%, respectively. Here, the absolute errors are minimal as well. For average waiting times from ICU to ward and ED to ICU, although the relative errors are above 10%, the absolute errors are not significant (0.02 and 0.31 hours only). Similarly, the relative errors for standard deviation of waiting times are extremely small (around 1%) except the waiting from ED to ICU. Although waiting from ED to ICU error is non-trivial, considering the error level in most prevailing queueing models, this level of error is deemed acceptable. Thus, we claim that the procedure can result in an acceptable accuracy.

In addition to queueing time, the accuracy of unit utilization is also defined and investigated. From Table 5.23, we claim that high accuracy in utilization estimation is obtained, all well below 5%.

$$\delta_{\rho_i} = \frac{|\widetilde{\rho}^{proc}_i - \widetilde{\rho}^{sim}_i|}{\widetilde{\rho}^{sim}_i} \cdot 100\%, \quad i = \mathrm{ed}, \mathrm{icu}, \mathrm{wd},$$

$$\Delta_{\rho_i} = |\widetilde{\rho}^{proc}_i - \widetilde{\rho}^{sim}_i|, \quad i = \mathrm{ed}, \mathrm{icu}, \mathrm{wd}.$$

Finally, the accuracy of unit full probability is studied. It is shown in Table 5.24 that high accuracy is obtained for the estimates. Note that although the relative error

Table 5.22: Accuracy of queueing time

$\delta_{\widetilde{q}_{\mathrm{icu,wd}}}(\%)$	$\Delta_{\widetilde{q}_{\mathrm{icu,wd}}}$ (hour)
11.27641	0.02626
$\delta_{\widetilde{q}_{\mathrm{ed,wd}}}(\%)$	$\Delta_{\widetilde{q}_{\mathrm{ed,wd}}}$ (hour)
1.25825	0.00444
$\delta_{\widetilde{q}_{\mathrm{ed,icu}}}(\%)$	$\Delta_{\widetilde{q}_{\mathrm{ed,icu}}}$ (hour)
13.86533	0.30955
$\delta_{\widetilde{q}_{\mathrm{wd,icu}}}(\%)$	$\Delta_{\widetilde{q}_{\mathrm{wd,icu}}}$ (hour)
5.66794	0.02217

(a) Average queueing time

$\delta_{\sigma_{\mathrm{icu,wd}}}$ (%)	$\Delta_{\sigma_{\mathrm{icu,wd}}}$ (hour)
0.17735	0.00046
$\delta_{\sigma_{\mathrm{ed,wd}}}$ (%)	$\Delta_{\sigma_{\mathrm{ed,wd}}}$ (hour)
1.01514	0.00699
$\delta_{\sigma_{\mathrm{ed,icu}}}$ (%)	$\Delta_{\sigma_{\mathrm{ed,icu}}}$ (hour)
17.39347	1.26984
$\delta_{\sigma_{\mathrm{wd,icu}}}$ (%)	$\Delta_{\sigma_{\mathrm{wd,icu}}}$ (hour)
1.75259	0.02605

(b) Standard deviation of queueing time

Table 5.23: Accuracy of unit utilization

$\delta_{\rho_{\text{ed}}}(\%)$	$\Delta_{\rho_{\text{ed}}}$
3.48217	0.02657
$\delta_{\rho_{\text{wd}}}(\%)$	$\Delta_{\rho_{\text{wd}}}$
4.77911	0.04393
$\delta_{\rho_{\text{icu}}}(\%)$	$\Delta_{\rho_{\text{icu}}}$
4.99153	0.04305

for π_{ed} is more than 10%, it is mainly due to small values as the absolute error is only 0.008.

$$\delta_{\pi_i} = \frac{|\pi_i^{proc} - \pi_i^{sim}|}{\pi_i^{sim}} \cdot 100\%, \quad i = \text{ed}, \text{icu}, \text{wd},$$

$$\Delta_{\pi_i} = |\pi_i^{proc} - \pi_i^{sim}|, \quad i = \text{ed}, \text{icu}, \text{wd}.$$

Table 5.24: Accuracy of unit full probability

$\delta_{\pi_{\text{ed}}}(\%)$	$\Delta_{\pi_{\text{ed}}}$
11.88424	0.00803
$\delta_{\pi_{\text{wd}}}(\%)$	$\Delta_{\pi_{\text{wd}}}$
0.61925	0.00246
$\delta_{\pi_{\text{icu}}}(\%)$	$\Delta_{\pi_{\text{icu}}}$
2.25372	0.00185

Therefore, we conclude that the proposed queueing network model based iterative

procedure can provide close estimates of system performance with acceptable accuracy.

Remark 5.5 The errors in performance estimation of the procedure mainly come from the following sources: First, only the first two moments are used to characterize the arrivals and services. Secondly, the G/G/c model using Kingman's formula is an approximation. □

5.5.3 Computation Efficiency

The introduced method is computationally efficient as well. Using Matlab on an Intel core i5-3340, 3.1 GHZ CPU, 12 GB memory computer, when the hospital capacity is about 800 beds in total, it takes 6 to 20 seconds to converge. When the hospital size is increased to 2000 beds, 1 to 3 minutes are typical to obtain the solutions. Beyond this capacity, the computation time grows substantially.

Although roughly 2000 beds is well above most hospital capacity, it is possible to go beyond and still be computationally efficient. Instead of a direct application of the iteration process, an approximation method can be pursued. Since high utilization is common in many large sized hospitals, the actual number of beds that is available to admit transfers becomes very limited. For instance, assume unit i, $i = \text{ed}, \text{wd}, \text{icu}$, has total capacity c_i, but $c_{occupy,i}$ beds are always occupied by patients. Then the number of available beds for admission/transfer will be $c_i - c_{\text{occupy},i}$, which can be used as the inputs for the analysis, i.e., input $c_i - c_{\text{occupy,i}}$ as the capacity of unit i. Note that the resulting queueing times and unit full probabilities can be directly used, but the utilizations need to be updated by including the unavailable beds, i.e., $[\rho_i(c_i - c_{\text{occupy},i}) + c_{\text{occupy},i}]/c_i$ will be the utilization of unit i. Therefore, even for hospitals with very large size, the model

and method introduced are still applicable through approximation.

5.6 Discussions

With a complete mathematical model describing patient transitions, next we seek to find some managerial insights in ways to reduce waiting times. To improve patient transitions and reduce delays, the first two improvement efforts should focus on the following [100]: Improve the service process by increasing the service capacity or synchronizing the capacity and patient arrival patterns; Improve the arrival process by altering the arriving patterns of patients to better align the capacity and the demand. Thus, in this section, using the complete model that captures patient transition behavior, we investigate system properties to provide some guidance in reducing patient wait times by focusing on bed capacity, admissions of external arrivals, and variabilities of service and arrival processes.

5.6.1 Bed Capacity

For hospitals with excessive waiting times, the most straightforward approach is to allocate more hospital beds. However, not only does adding a bed to different unit have varying impacts on waiting time reduction, increasing bed capacity can actually worsen the blockages in some cases. Below such impacts are discussed in more details.

Increasing ED Capacity

As ED capacity increases, more patients flow into the ED (admitted to the ED) and more patients need to flow out from the ED (transfer from ED to other units). Thus the waiting times of patients transferring out from the ED increase. The waiting times

of transitions between ICU and ward show varying patterns depending on the system parameters. In general, when ED bed occupancy is high (roughly $> 80 - 85\%$), the ED may become a source of congestion in the hospital. As external arrival is the only source of admission for ED, increasing ED capacity leads to more patients flowing into the whole system. Thus, as the total number of patients in the hospital increases, the waiting times among all units become longer. As one can observe in Figure 5.18, the waiting times increase slightly.

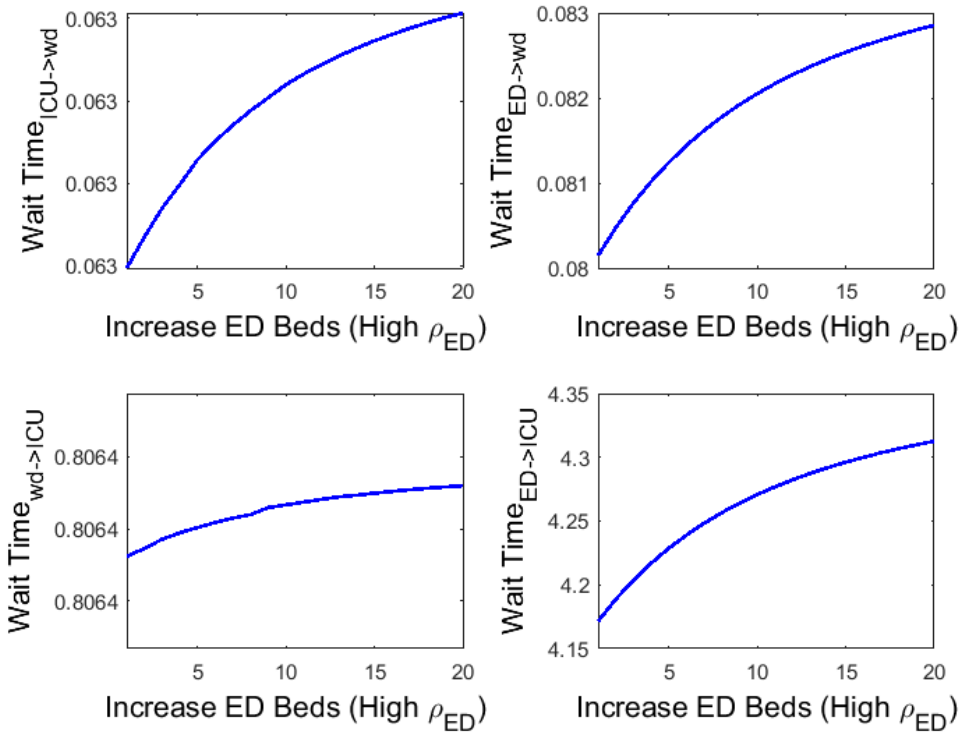

Figure 5.18: Increasing ED capacity: High utilization case

On the other hand, when ED bed occupancy is low, increasing ED capacity does not result in such a high inflow of patients into the system. Thus the transition of patients between ICU and ward remains mostly unaffected. Even though more patients need to

be transferred from ED to these units, as ED patients have the lowest priority, their impact is not significant. As shown in Figure 5.19, the waiting times in ICU and ward do remain the same.

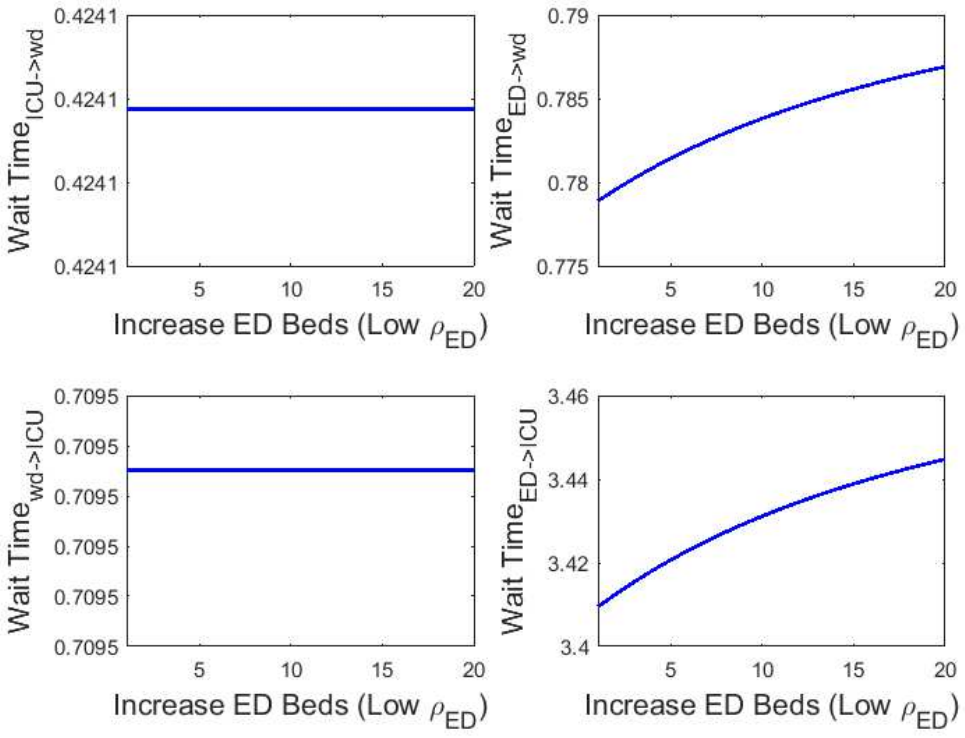

Figure 5.19: Increasing ED capacity: Low utilization case

Increasing Ward Capacity

When more beds are allocated to the ward, patients need to wait less time to transfer to ward from other units. Similarly to the ED case, high ward capacity implies ward could become a source of bottleneck of the system. However unlike the ED, ward has multiple sources of admission where external arrivals have the lowest priority. Thus the increased ward beds are mainly used to admit the internal patients transiting from other units.

When ward utilization is high (e.g., $> 80 - 85\%$), more internal patients are already waiting to be transferred, thus increasing ward beds do not lead to admission of more external arrivals. Since the total number of patients in the system does not change significantly although the number of ward beds increases, all waiting times can be reduced slightly. This is validated in Figure 5.20.

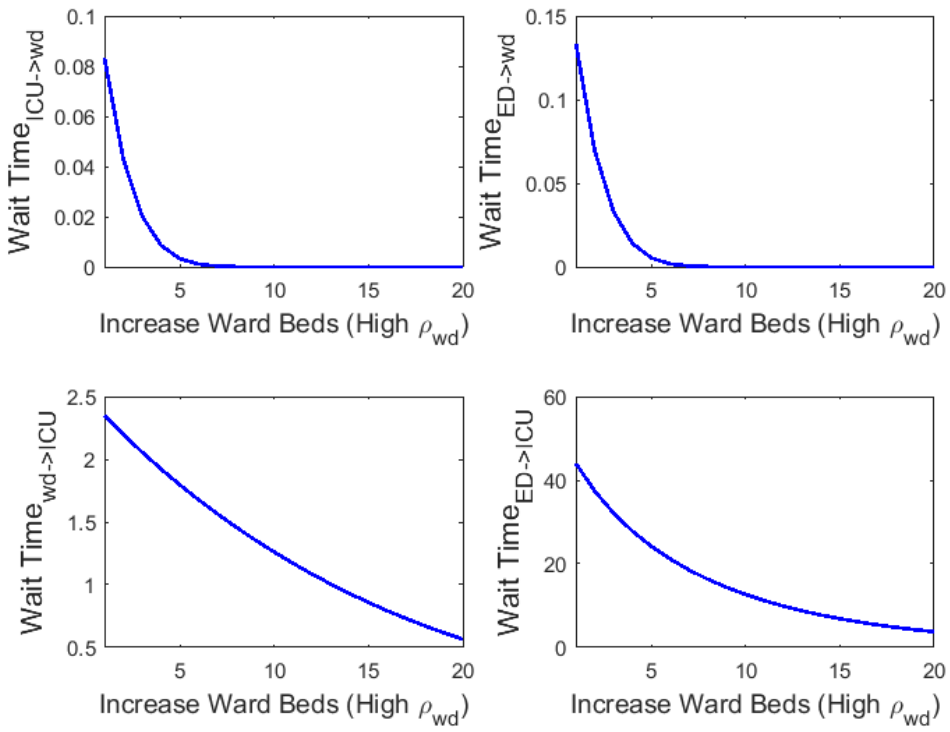

Figure 5.20: Increasing ward capacity: High utilization case

However, when ward bed utilization is low, increasing ward beds leads to admission of more external arrivals so that more ward patients occupy the department. Thus patients who need to transfer to other units from ward could experience a little longer waiting time, as shown in Figure 5.21.

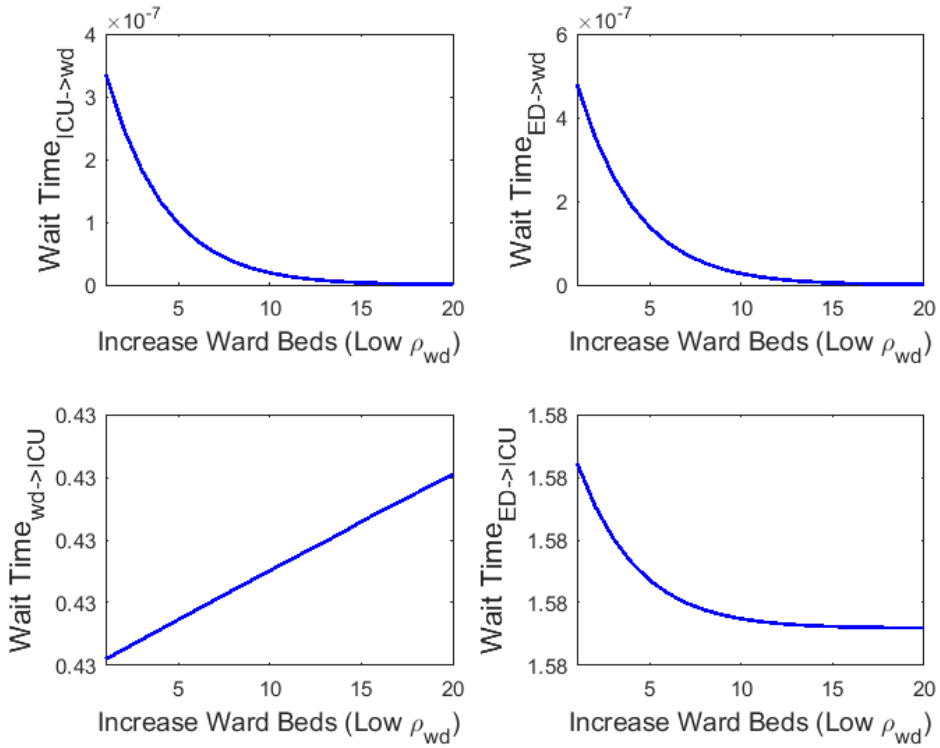

Figure 5.21: Increasing ward capacity: Low utilization case

Increasing ICU Capacity

In contrast to ED and Ward, ICU does not have direct external arrivals and thus is more dependent on the other two units. Obviously, as more beds are allocated to ICU, waiting times of the patients transiting into ICU decrease. However, other waiting times do not show a clear pattern and are mainly dependent on two factors: bed utilizations in ward and ICU.

When bed occupancy of ward is higher than that of ICU, increasing ICU bed capacity may worsen the waiting times of patients transiting from ICU to ward. As shown in

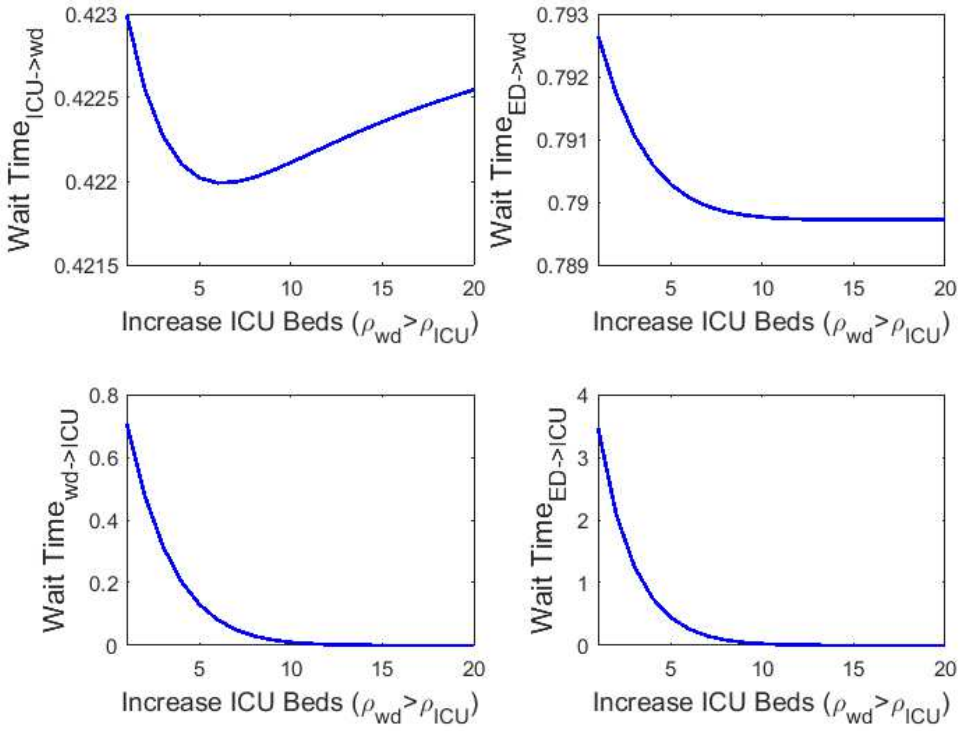

Figure 5.22: Increasing ICU capacity: High ward utilization case

Figure 5.22, waiting times are non-monotonic, where adding a few ICU beds slightly

reduces the waiting times when the bed number is small, but adding too much results in longer waiting times. When ICU bed capacity is very low, adding more beds can accept more patients where majority of them will be coming from ward (as ward has higher priority than ED). Hence, ward occupancy will decrease which leads to less waiting for ICU patients to transfer to ward. However, as more ICU beds are added, ED patients will be accepted as well, thus the ward availability will not be affected as much. While the discharge rate from the ward remains similar, as adding ICU beds implies more patients requesting to transfer from ICU to ward, ward admission requests increase. Thus, as more ICU beds are added, the waiting time from ICU to ward increases.

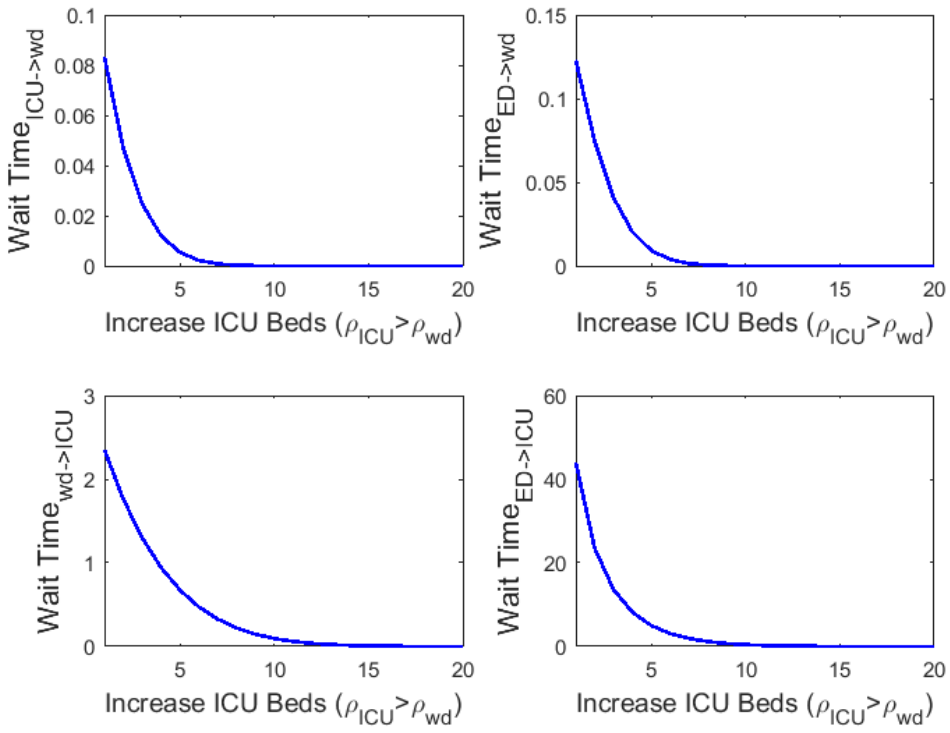

Figure 5.23: Increasing ICU capacity: High ICU utilization case

On the other hand, when ICU bed occupancy is higher than ward, increasing ICU

beds actually helps in resolving the blockings, thus the waiting time of transitions from ICU to ward decreases (see Figure 5.23).

For highly overcrowded hospitals that have over 80-85% bed occupancy level for ED and ward, efforts should be focused on increasing ward beds. With overcrowded ED being one of the major issues in many hospitals, increasing ED beds may seem like the solution as the waiting time to enter ED or the rate of leaving without being seen can drop initially. However, this may be merely due to the shift of blockage from external admissions to internal units. Thus, to solve the overcrowded ED problems, hospitals should be considered as an integrated system where one unit's capacity impacts other units.

5.6.2 External Admission Rate

For hospitals with excessive waiting times, although increasing bed capacity can help reduce waiting times, cost of extra hospital beds may be an economic burden and limited floor space may also restrict the expansion. Thus, instead of increasing the capacity, decreasing the arrivals can be an alternative approach which can be achieved by referring patients to other hospitals or controlling admission scheduling policy. As two types of external admissions, emergency (to ED) and elective (to ward) admissions, are considered in this model, decreasing which admission brings the most reduction in blocking is studied. By reducing the same number of patient admissions per day, the impacts of reducing ED admission (Figure 5.24) and reducing ward admission (Figure 5.25) are compared.

As one can see, decreasing ward admission brings steeper reduction in waiting times where the difference is the greatest for waiting times to enter ward. Admitting one less patient per day can impact patient waiting time significantly, where the longer

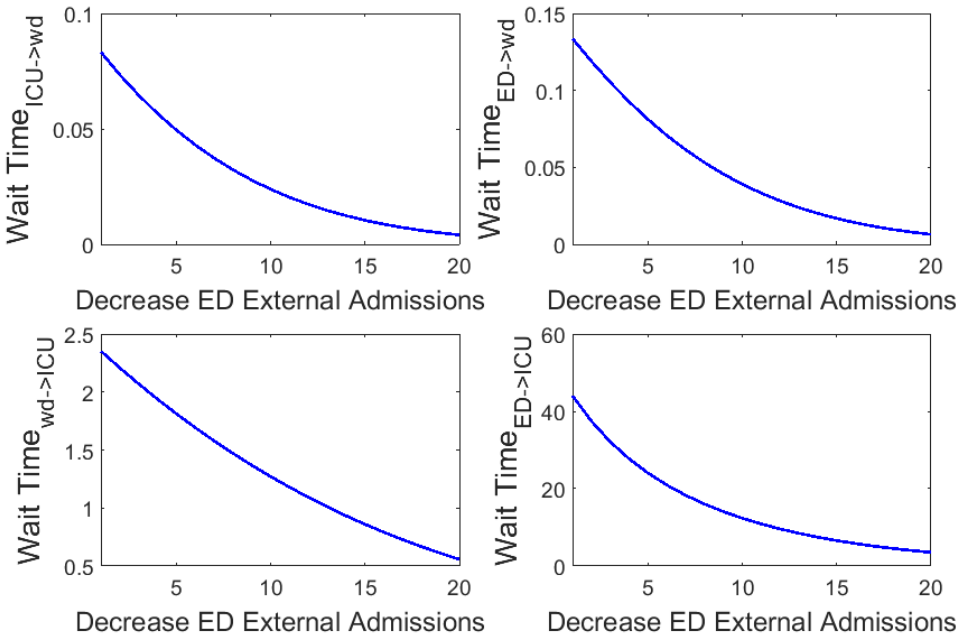

Figure 5.24: Reducing ED admission

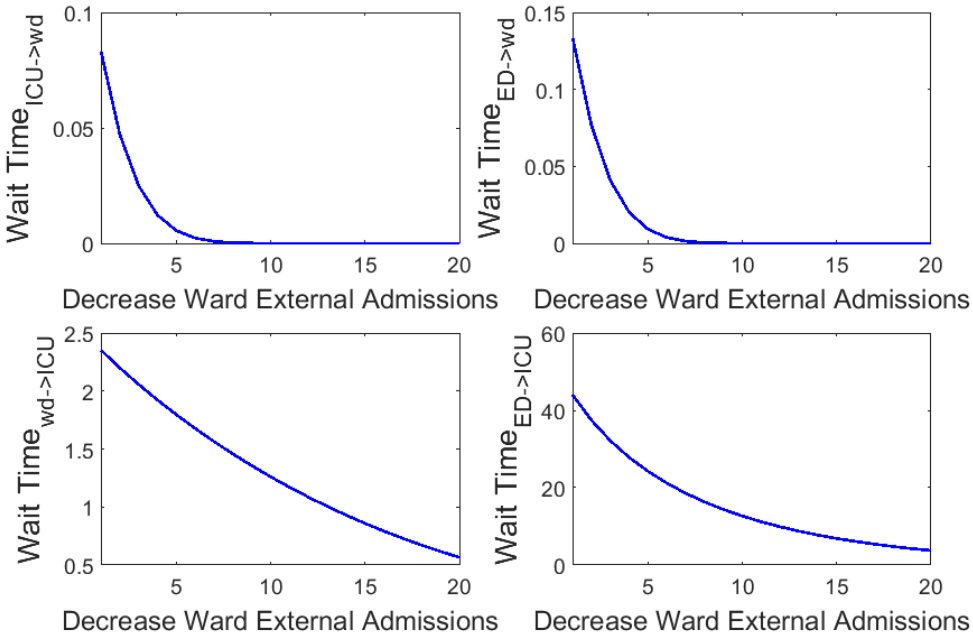

Figure 5.25: Reducing ward admission

the waiting time is, the larger the reduction is. Thus, when external admissions need to be controlled, reducing ward admission is more effective in resolving the blockings. Moreover, controlling ward admissions is more achievable than restricting ED admissions as ED patients' arrivals are unpredictable and uncontrollable.

5.6.3 Variability

As an effort to improve the service or arrival process, controlling the variability to better align the service capacity and patient arrivals can be pursued as well. Reducing variability does have more limiting aspects as external arrivals are difficult to control due to emergency admissions and service times are hard to control due to their dependence on individual patient's health status. However, understanding how variability impacts the system is still critical in hospital operation management. Thus, extensive numerical experiments have been carried out to investigate the impact of arrival and service time variabilities on unit waiting times. Based on the experiments, we summarize the observations as follows.

Variability of External Arrivals

It is shown that all waiting times are monotonically increasing with respect to both the CVs of external admissions to ED and ward, i.e., $CV_{\text{ext,ed}}$ and $CV_{\text{ext,wd}}$. However, their impacts on the waiting times differ.

- For the waiting times to transfer out from the ED, $\widetilde{q}_{\text{ed,wd}}$ and $\widetilde{q}_{\text{ed,icu}}$, we notice that $CV_{\text{ext,ed}}$ has larger impact than $CV_{\text{ext,wd}}$. This could be due to that ED has only one source of arrival which is the external admission to ED, thus the ED wait times are mainly impacted by the CV of ED's external admissions.

- For the waiting to transit from ward to ICU, as ward has multiple sources of arrivals, both the external admissions to ED and ward can impact the waiting time. Hence, the degree of impact depends on the system settings. As a unit's discharge variability can be expressed as:

$$(CV_{\text{dc}})^2 = \rho^2(CV_{\text{service}})^2 + (1-\rho^2)(CV_{\text{arr}})^2,$$

  it follows that as ρ decreases, the impact of CV_{arr} increases. Thus, when $\rho_{\text{wd}} < \rho_{\text{ed}}$, $CV_{\text{arr}_{\text{wd}}}$ has higher impact, which leads to $CV_{\text{ext,wd}}$ having higher impact; while when $\rho_{\text{wd}} > \rho_{\text{ed}}$, $CV_{\text{arr}_{\text{ed}}}$, which implies $CV_{\text{ext,ed}}$, impacts the waiting time more severely.

- However, the impact of $CV_{\text{ext,ed}}$ and $CV_{\text{ext,wd}}$ on the waiting time from ICU to ward does not show a clear pattern and is mainly dependent on system parameters.

Although reducing $CV_{\text{ext,ed}}$ could be helpful in reducing the waiting times from ED, controlling emergency admissions may sometimes be impossible to achieve. Moreover, for most hospitals that suffer from high ED occupancy rates, if ward is less utilized than ED, controlling ward external admissions could be more effective in reducing blockings.

Variability of Service Times

Similarly, we observe that all waiting times are monotonically increasing with respect to the service time CVs of both the ED and ward, i.e., CV_{ed} and CV_{wd}. Whereas for ICU's service variability, all waiting times except $\widetilde{q}_{\text{ed,wd}}$ are monotonically increasing with respect to CV_{icu}.

For $\widetilde{q}_{\text{ed,wd}}$, when both ρ_{wd} and ρ_{icu} are high (roughly $> 80\%$), counter-intuitively, it is found that $\widetilde{q}_{\text{ed,wd}}$ is a decreasing function of CV_{icu}. This could mainly be due to the fact that when ward beds are highly utilized, in most cases, only higher priority

(ICU) patients are able to be admitted to ward, thus the lower priority (ED) patients' waiting time $\widetilde{q}_{\text{ed,wd}}$ tends to be high. Now, as CV_{icu} increases, this directly increases the discharge variability of ICU, where for high ρ_{icu} case, the impact of CV_{icu} on discharge variability is larger. This in turn increases the transition variability from ICU to ward. Hence, as the arrivals of higher priority (ICU) patients have large variability, the lower priority (ED) patients increasingly have more chance to be admitted to the ward. Thus when both ward and ICU beds are highly utilized, increasing CV_{icu} can help reduce the waiting times from ED to ward.

Now, comparing the degree of impact each service time CV has on waiting times, the following pattern is observed. For waiting time to enter a destination unit from an originating unit, the service time CV of the destination unit has higher impact than that of the originating unit. The other remaining unit has insignificant impact on waiting time. Thus the service time CVs' impact can be ordered as follows:

$$\begin{aligned}
&\text{Impact on } \widetilde{q}_{\text{icu,wd}}, && CV_{\text{wd}} > CV_{\text{icu}} > CV_{\text{ed}},\\
&\text{Impact on } \widetilde{q}_{\text{wd,icu}}, && CV_{\text{icu}} > CV_{\text{wd}} > CV_{\text{ed}},\\
&\text{Impact on } \widetilde{q}_{\text{ed,icu}}, && CV_{\text{icu}} > CV_{\text{ed}} > CV_{\text{wd}},\\
&\text{Impact on } \widetilde{q}_{\text{ed,wd}}, && CV_{\text{wd}} > CV_{\text{ed}} > CV_{\text{icu}}.
\end{aligned}$$

Thus, although controlling the variability of service times may be infeasible in some cases, whenever possible, efforts should be focused on reducing the variability of the unit to where the longest waiting times are formed.

5.7 Conclusions

This chapter introduces a queueing network model to study patient transitions between ED, ICU, and general ward units within a hospital. An iterative approach is presented to evaluate the transition waiting time, department utilization, and probabilities of full occupancy. Furthermore, with the developed model, system properties are investigated to seek effective ways of reducing system delays. Specifically, the impacts of bed capacity, external admission rates, and variabilities of the system are studied. It is shown that increasing ward capacity or controlling the elective admissions is often times more effective in improving the transitions. Such a method provides a quantitative tool for managing transition processes.

Chapter 6

Optimal TJR Postoperative Care Management

6.1 Introduction

In this chapter, we develop a finite-horizon discrete-time Markov decision process (MDP) to model the postdischarge intervention process following TJR surgeries. Specifically, we dynamically model the post-TJR process by directly incorporating the readmission risk and penalty in the model, and considering the varying effectiveness of interventions depending on where the patient is located at (care facility type). The MDP model is introduced in Section 6.2, and a number of structural properties are presented in Section 6.3. Section 6.4 provides a detailed analysis of the data used to model the postdischarge intervention procedure and estimate the system parameters. In Section 6.5, we present numerical results illustrating the optimal policy and conduct sensitivity analysis. Finally, conclusions are formulated in Section 6.6.

6.2 Markov Decision Process Model Formulation

The optimal post-TJR intervention problem is formulated as a finite-horizon discrete-time Markov decision process [101] where the state transition diagram is shown in Figure

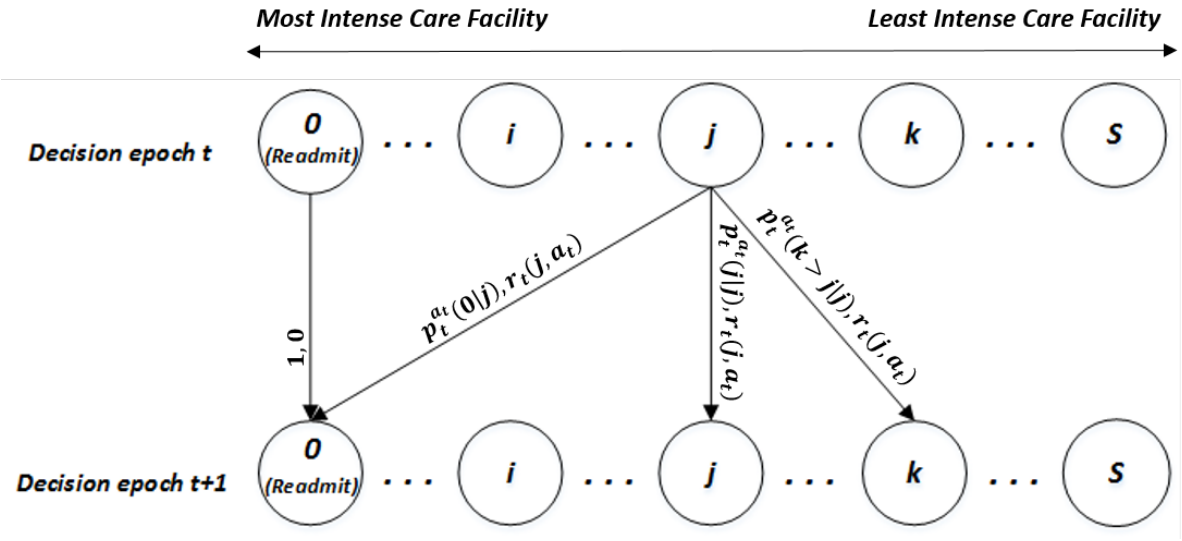

Figure 6.26: MDP model structure: system states and possible transitions

6.26. In general, there exist varying levels of care facilities (e.g., skilled nursing facility (SNF), rehabilitation facility, home service, self care, etc.) TJR patients can reside in depending on the patient's health status and insurance eligibility. In most cases, patients will not stay at the same care facility during the entire phase of recovery; a stay at one care facility may last from a few days to a few weeks. Thus, expect for self care, patients will de-elevate to a lower level care facility as they recover throughout time.

In our model, starting from the initial discharged care facility, patients progress through care facilities (de-elevate to a lower facility or remain at the same facility) as time passes. At each discretized time point, health professionals have the choice of intervening the patient or not, and as a result, the patient's readmission risk and de-elevation probabilities are affected accordingly. Note that intervention here refers to additional supports or rehabilitative programs rather than direct treatments or medications. The objective of the model is to maximize expected total benefit (or minimize expected total cost) for the TJR patient. The MDP is characterized by a state space, an action set,

decision epochs, costs, and transition probabilities where the details of each component are as follows.

• Decision epoch: $t \in \{0, 1, 2, ..., T\}$ where time t represents the time passed since discharge. One decision epoch needs not be equivalent to one day; it can represent multiple number of days (one week, 10 days, etc.). If one decision epoch represents multiple days (m days), planning horizon N needs to be set as a multiple of m (i.e., $N = T * m$) to ensure evenly spaced decision epochs. Thus, decision epoch t corresponds to $m * t$ days from discharge. For the boundary points, decision epoch $t = 0$ corresponds to the time of discharge from the hospital, and decision epoch $t = T$ corresponds to the end of the planning horizon; thus no decision is made at this point and is included to evaluate whether a patient has been readmitted or not.

• State: $s_t \in \{0, 1, ..., S\}$ represents the state of the system at time t where S is the care facility options for TJR patients, i.e., $s_t \in \{1, ..., S\}$ represents the care facility where the patient is located in at decision epoch t and $s_t = 0$ represents the readmitted state. We assume a complete ordering (descending order) of the care facility states $\{1, ..., S\}$, thus 1 represents the care facility with the most intense level of care and S the care facility with the least intense level of care.

• State space: $\mathbb{S} = \{0, 1, ..., S\}$.

• Action: $a_t(s_t) \in \{\text{Wait (W), Intervene (I)}\}$ represents the action taken in state s_t at decision epoch t. At each decision epoch, the decision maker either gives intervention to the patient (I) or waits and does nothing (W).

• Action space: $\mathbb{A} = \{\text{Wait (W), Intervene (I)}\}$.

• Transition Probability: $p_t^{a_t}(s'|s_t)$ denotes the probability that the patient will be in state $s' \in \mathbb{S}$ at decision epoch $t + 1$, given that he/she is in state s_t at decision epoch t and the chosen action is $a_t \in \mathbb{A}$. We assume transitions occur immediately at the start

of the next decision epoch $t+1$, i.e., if patient is in care facility i at the beginning of decision epoch t and action a_t is chosen, patient will remain in care facility i until the end of decision epoch t, and then transit immediately to care facility j at the start of the next decision epoch $t+1$ with probability $p_t^{a_t}(j|i)$.

For $t = 0, 1, ..., T-1$, there are 3 types of possible transitions.

(i) $p_t^{a_t}(0|s_t)$: Probability of patient getting readmitted at decision epoch $t+1$.

Remark 6.1 For $s_t = 0$, we define $p_t^{a_t}(0|0) = 1$, i.e., readmitted state is an absorbing state. □

(ii) $p_t^{a_t}(s' > s_t|s_t)$: Probability of patient de-elevating to a lower level care facility s' at decision epoch $t+1$.

(iii) $p_t^{a_t}(s' = s_t|s_t)$: Probability of patient remaining at the same care facility at decision epoch $t+1$.

Remark 6.2 Note that we assume elevation to a higher level care facility never occurs. Careful tracing of patient records and consultation with healthcare professionals at the collaborating hospital reveal that elevation involves insurance issues and thus typically happens very rarely. Once patient's initial disposition location is decided, it is difficult to elevate the patient to a higher level care facility. Thus, without loss of generality, we assume no elevation happens. □

• Reward: $r_t(s_t, a_t)$ denotes the total expected reward accrued at decision epoch t, when the patient is in state s_t and action a_t is chosen.

Remark 6.3 For $s_t = 0$, we define $r_t(0, a_t) = 0$, i.e., once readmitted, patient quits the process and receives no reward in future decision epochs. □

For $a_t = I$, $r_t(s_t, I) = -$(Intervention Cost + Care Facility Cost) $= -(C_I + C_{CF}(s_t))$.

For $a_t = W$, $r_t(s_t, W) = -$Care Facility Cost $= -C_{CF}(s_t)$.

Here, C_I denotes the one-time cost of providing intervention to the patient. We assume at most one intervention is given to the patient during one decision epoch. $C_{CF}(s_t)$ represents the expenses accrued for the patient staying at care facility s_t during decision epoch t. The total expenses accrued during a decision epoch are assumed to be incurred as a one-time lump-sum cost at the beginning of that epoch.

• $v_t(s_t)$: Maximum total expected reward when patient is in state s_t at decision epoch t. Since the state and action space are finite, the existence of a Markovian, deterministic policy is guaranteed where the optimal solution can be obtained by solving the following set of recursive equations [101]:

$$v_t(s_t) = max\Big\{r_t(s_t, a_t) + \sum_{s' \in \mathbb{S}} p_t^{a_t}(s'|s_t) v_{t+1}(s')\Big\}, t = 0, 1, ..., T-1. \tag{6.1}$$

For $t = T$, the boundary condition is given as follows:

$$v_T(s_T) = r_T(s_T, I) = r_T(s_T, W) = \begin{cases} V_{\text{Nonreadmit}} & \text{for } s_T \in \{1, ..., S\}, \\ 0 & \text{for } s_T = 0. \end{cases} \tag{6.2}$$

The boundary condition implies we only consider whether a patient is readmitted or not during the planning horizon, i.e., the final care facility location of the patient is not in the scope of the decision making. Since readmission penalty only considers readmissions up to a fixed number of days from discharge (90 days for TJR), patient location after decision epoch T is not of interest to the decision maker. Thus, we define the reward incurred at decision epoch T as zero if the patient is readmitted and $V_{\text{Nonreadmit}} \geq 0$ if the patient is not readmitted. In this sense, $V_{\text{Nonreadmit}}$ denotes the value of a nonreadmitted patient in terms of cost. Specifically, we use the average readmission penalty per person to assume the relative value of a nonreadmitted patient ($V_{\text{Nonreadmit}}$).

Remark 6.4 Note that the actual way CMS calculates readmission penalty is not according to per readmitted patient. CMS measures a hospital's performance by calculating the excess readmission ratio (ERR), which is the ratio of predicted-to-expected readmissions. This ERR links directly to the hospital's readmission penalty, the greater the rate of excess readmissions, the higher the penalty. Here we consider such a per person definition for analysis purpose, where it can be obtained by dividing the total readmission penalty for a given period by the total number of readmissions occurred in the period. □

6.3 Structural Properties

In this section, we discuss the structure of the post-TJR intervention model introduced in Section 6.2 and prove several structural properties that provide foundations for developing the optimal policy. Before we examine the underlying structural properties of the MDP model, we first introduce assumptions that are used throughout the section.

Assumption 6.1. Care facility cost $C_{CF}(s_t)$ depends only on the care facility type but not on time. That is, care facility cost per day remains constant over the course of the patient stay at the care facility, i.e., daily cost does not depend on how long the patient has been staying at the facility. Regarding the care facility type, $C_{CF}(s_t)$ is nonincreasing in s_t for all t, which implies cost does not increase as the level of care facility decreases.

Assumption 6.2. Intervention cost C_I does not depend on care facility or days since discharge. That is, patients in different care facilities receive the same intervention and the intervention cost does not change over time.

Assumption 6.3. For each action, readmission probability $p_t^{a_t}(0|s_t)$ is nonincreasing in

s_t for all t, and in t. This implies that the readmission risk does not increase as the patient is de-elevated to a lower level care facility or longer days have elapsed since discharge.

Assumption 6.4. For each action, the reward function $r_t(s_t, a_t)$ is nondecreasing in $s_t \in \{1, ..., S\}$ for all t, and in t. This implies that the total expected reward does not decrease as the patient is de-elevated to a lower level care facility or longer days have elapsed since discharge.

Assumption 6.5. For each action, the cumulative probability of de-elevating to a lower level care facility increases as more days have elapsed since discharge. The longer the days since discharge, the more probable that the patient will be de-elevated to a lower level care facility. In other words, a patient that has been discharged earlier is more likely to be de-elevated to a lower level care facility.

Assumptions 6.1 and 6.2 are straightforward. For the analysis to be tractable, we assume a constant daily care facility usage cost for all facility types. Also, in accordance to intuition, facility cost becomes more expensive as the level of care provided at the facility increases (more intense level of care). Hence, $C_{CF}(s)$ is nonincreasing in s. Assumption 6.2 holds true if we consider a single type of intervention throughout the decision making horizon. Note that intervention refers to additional supports or rehabilitative programs provided by the hospital rather than direct treatments or medications. Thus, if we provide the same type of intervention during the entire planning horizon, we can assume the intervention cost is constant over time.

Supporting evidence for Assumption 6.3 can be found in related medical literature. First, regarding the assumption of nonincreasing readmission probability in time, it is identified in prevailing literature that majority of the readmissions occur in the early

weeks from discharge: according to reference [102], of a total of 591 unplanned readmissions that occurred within 90 days of discharge, 348 (58.9%) occurred within the first 30 days of discharge. A more detailed distribution of readmission day is presented in [103] where the distribution of emergency readmissions as a function of number of weeks after discharge is generated using 7547 data points. The readmission numbers are considerably higher in the first two weeks after discharge and reach a steady state by about four weeks after discharge. Paper [104] analyzes 160 patient records by discretizing the time from discharge by a week, and identifies that 35% occur in the first week, 25.5% in the second and third week, and 14% in the fourth week. Hence, when the time epochs are set appropriately, it is safe to assume readmission probability is nonincreasing in time.

Regarding Assumption 6.3 where readmission probability is nonincreasing in s_t, there are limited studies specifically examining the readmission rates by discharge disposition. Of those, paper [105] demonstrates that patients discharged home with health services had a significantly lower readmission rate compared to those discharged to inpatient rehab facilities. Paper [106] compares 90-day readmission rates in patients discharged home with health services versus those discharged to a SNF, and identifies a higher risk for readmission in patients discharged to a SNF. Also, it is identified in [102] that discharge to a SNF or rehab center rather than home is associated with an increased likelihood of readmission. Thus, based on the supporting literatures, we assume less intense level of care facility is associated with lower readmission rates.

Assumption 6.4 immediately follows from Assumptions 6.1 and 6.2. The reward function consists of intervention cost and care facility cost, where due to Assumptions 6.1 and 6.2, C_I does not depend on s_t and t, and C_{CF} is nonincreasing in s_t and non-dependent on t. Thus, the resulting sum of the costs is nonincreasing in s_t and non-dependent on t, which implies the reward function is nondecreasing in s_t and t.

Assumption 6.5 can be interpreted as IFR (increasing failure rate [107]) property which can be expressed as $\sum_{s'=k}^{S} p_t^{a_t}(s'|j) \leq \sum_{s'=k}^{S} p_{t+1}^{a_{t+1}}(s'|j)$ for $j, k \in \mathbb{S}$. Similar approach can be found in papers dealing with clinical decision-making problems [90, 108].

Based on these assumptions, we obtain the following propositions.

Proposition 6.1 *Given Assumptions 6.1-6.5, $v_t(s_t)$ is nondecreasing in s_t for $s_t \in \{1, 2, ..., S\}$ and $t \in \{0, 1, ..., T-1\}$.*

Proposition 6.1 implies the total expected reward does not decrease as a patient is de-elevated to a lower level care facility. The proof of this proposition is given in Theorem 4.7.3 in [101] and hence is omitted.

Proposition 6.2 *Given Assumptions 6.1-6.5, $v_t(s_t)$ is nondecreasing in t for all $s_t \in \mathbb{S}$.*

Proof: See Appendix D. ■

Proposition 6.2 implies $v_{t+1}(s) \geq v_t(s)$, which can be interpreted as follows: the total expected reward does not decrease as longer days have elapsed since discharge. That is, as patients are not readmitted for longer days, their expected reward will never decrease.

Thus, from Propositions 6.1 and 6.2, $v_t(s_t)$ is nondecreasing in s_t for all t, and in t; i.e., the total expected reward does not decrease as patients are de-elevated to a lower level care facility or longer days have elapsed since discharge. Given such structural properties, we next analyze the EHR data to define probable intervention process and derive optimal policies.

6.4 Estimation of Model Parameters

The clinical data (EHR) used throughout this section come from St. Mary's Hospital (SMH) of SSM Health. In the rest of this section, we first investigate the current intervention process that is conducted in practice at the hospital. Based on the findings, we define probable intervention process and estimate the parameter values for the MDP model.

6.4.1 Postoperative Intervention Process

To define probable interventions that reduce a patient's readmission risk, a detailed analysis of current post-surgery care process at the collaborating hospital is carried out. When a TJR patient is discharged, based on the patient's progress during hospitalization, availability of support at home, and medical complications, disposition is decided among three options: stay at a SNF; stay at home but receive home care visits provided by professional services (home service); and stay at home with no additional services (home or self care). Each discharge option has varying degree of care support where SNF is the most intense level care facility and self care is the least. Ideally, more severe patients are discharged to higher level care facilities where they receive the most appropriate level of treatment. Hence, treatment intensity and frequency differ across care facilities where higher level care facilities are associated with more intense treatment plans.

In addition to treatments, hospitals provide intervention or rehabilitative services to the patients where in general, all patients go through similar processes regardless of their care facility type. Specifically, patients mainly receive three types of intervention care processes. First, the orthopedic care coordinator nurse calls the patient within 24 hours after discharge to check on the recovery status. Phone call is generally conducted

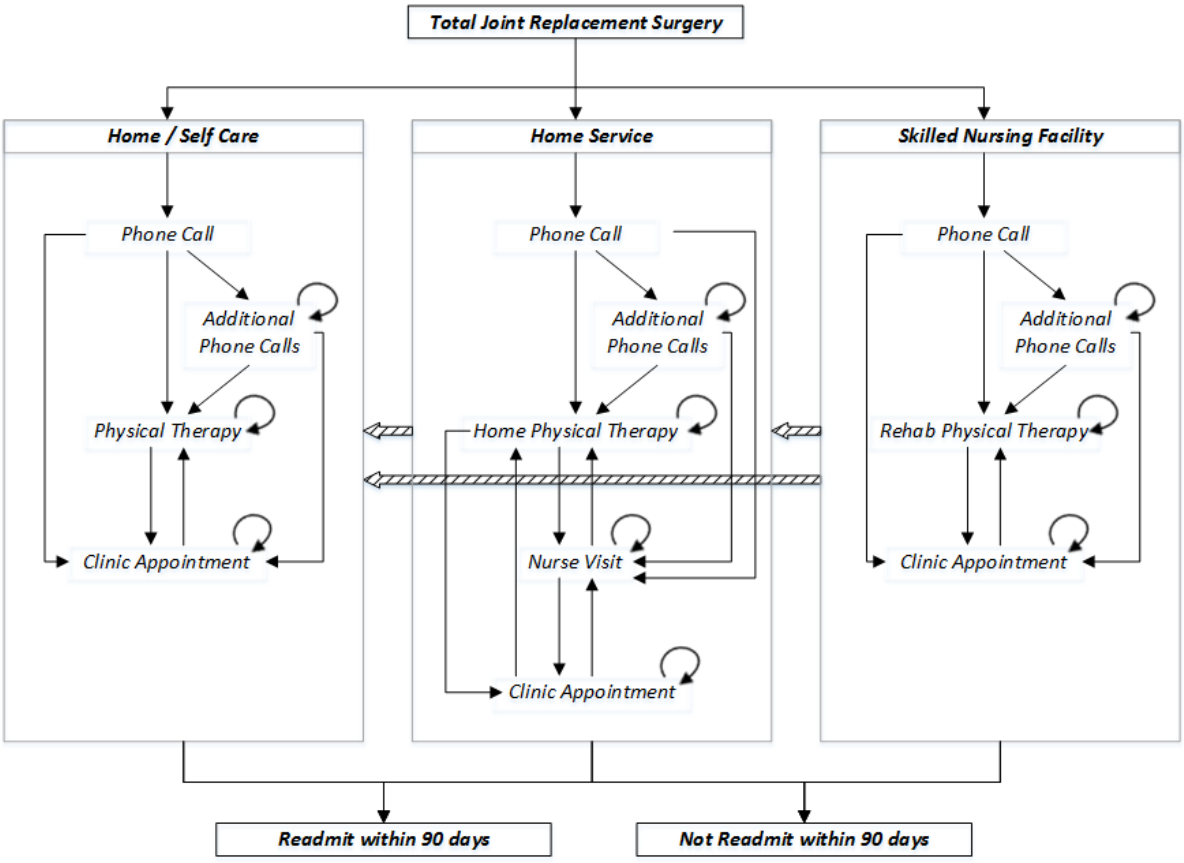

Figure 6.27: Postoperative intervention process at the collaborating hospital

once, but if the nurse feels that the patient condition is not stable, additional phone calls are conducted throughout the following days. In addition to phone calls, patients also receive PT and are scheduled for clinic visits on a regular basis. TJR patients are scheduled to visit the orthopedic physician in 2 weeks, 6 weeks, 1 month, 3 months and 1 year post discharge and are recommended to receive PT exercises on a weekly basis. An illustration of the post-surgery care process is shown in Figure 6.27. Note that PT may be received in different forms depending on where the patient is located at. SNF patients receive rehabilitative therapy at the built-in PT facility at SNF, home service patients receive PT by home care professionals (therapist, nurse) visiting their homes, and self care patients visit external physical therapy centers.

During this course of recovery, patients may be de-elevated to a lower level care facility (see the thick dashed arrows in Figure 6.27). Typically, this transition is facilitated due to the insurance limits imposed on the number of days patients can stay at a certain care facility. However, some patients may de-elevate earlier if they have recovered enough to safely transit to a lower level of care facility. Once a patient has been de-elevated, he/she will continue the intervention process at the new care facility. One thing to note is that hospital readmissions can occur any time during this course regardless of where the patient is located at, and once readmitted, the patient is classified as a readmitted case regardless of future trajectories. Thus, given a predetermined planning horizon, only if a patient has not been readmitted until the end of the planning horizon, that patient is classified as a non-readmitted case.

In summary, there are three levels of care facilities: SNF, HS (home service) and H (home), and three types of interventions: phone call, clinic visit and physical therapy. To derive general properties of such interventions, a total of 180 patient records obtained from the collaborating hospital are analyzed, of which 52 are readmitted cases and 128

are non-readmitted cases. As the specific type of interventions may vary among hospitals, in order to provide a general modeling scheme, we group all three types of interventions and analyze as a single one. Thus, we examine how the readmission probabilities and de-elevation probabilities change depending on whether the patient receives any type of intervention or not in a given time period. Regarding the time period, since SMH is interested in reducing readmission penalties from CMS, the planning horizon is set as 90 days from discharge. For time epochs, as intervention decisions are not evaluated on a daily basis and defining one time epoch as one day does not satisfy Assumption 6.3 (readmission probabilities do not show nonincreasing pattern), decision epoch is carefully determined. Specifically, 90 days are discretized into 9 equally sized bins, i.e., one time epoch corresponds to 10 days.

6.4.2 Readmission Risk and Transition Probability Estimation

Now, prior to estimating the transition probabilities, we first introduce an assumption regarding the effectiveness of interventions. We assume interventions result in a reduced readmission risk and increased chance of de-elevation for all care facilities, i.e., if a patient received intervention at time epoch t, the readmission risk at time epoch $t+1$ will be lower and de-elevation probability at time epoch $t+1$ will be higher than the patient who did not receive an intervention. Hence, the following assumption is introduced:

Assumption 6.6. Interventions reduce patient's readmission probability and increase the chance of de-elevation to a lower care facility, i.e., $p_t^I(0|s) \leq p_t^W(0|s)$ and $\sum_{s'=k}^{S} p_t^I(s'|s) \geq \sum_{s'=k}^{S} p_t^W(s'|s)$.

By analyzing the 180 patient records, transition probabilities for the model are estimated using maximum likelihood estimators (MLE). Specifically, the MLE of each transition probability is defined as the ratio of the number of transitions for a specific

transition over the total number of transitions that have a common exit state [108]. Here, a patient who received at least one intervention during time epoch t to $t+1$ is considered as having taken action $a = I$. On the other hand, if no intervention was given during time t to $t+1$, it is assumed action W has been taken.

First, regarding the data points corresponding to action "Intervene", majority of the data points (65%) come from patients located at home (self care). Thus, the number of data points for SNF or HS patients at certain time epochs is not sufficient to estimate the transition probabilities. Specifically, for the time epochs where the number of data points is less than 20, we estimate the readmission probabilities by the following equation:

$$p_t^I(0|s) = \begin{cases} 0 & \text{if } p_t^I(0|H) = 0, \\ p_{t-1}^I(0|s) - [p_{t-1}^I(0|H) - p_t^I(0|H)] & \text{otherwise.} \end{cases} \tag{6.3}$$

Equation (6.3) implies that the difference in readmission risk between previous and current time epochs for home patients is assumed to apply to other care facility patients as well. Thus, it assumes regardless of where the patient is located at, his/her readmission risk decreases in time by the same amount. An exception occurs when home patient's readmission probability is zero. In this case, we assume other care facility patients have zero readmission risk to ensure their readmission probability converges to zero as well. For the transitions within care facilities, the complement of readmission probability, i.e., $(1 - p_t^I(0|s))$, is distributed among the transition probabilities $p_t^I(s'|s)$ while maintaining the original ratio in the observed data points.

Regarding action "Wait", the transition probabilities for $a = W$ refer to the patient's baseline transition probabilities when only necessary treatments are provided without additional rehabilitative supports. To the best of our knowledge, there are no medical literatures that address TJR patients' baseline readmission risks since most hospitals

conduct some form of intervention processes on their patients. Therefore, we introduce two scenarios under which we assume different baseline transition probabilities, both of which are clinically meaningful.

Scenario 1: Constant baseline readmission risk in care facility level

In Scenario 1, we assume the most appropriate level of treatment is provided to the patients for each care facility type. Thus, the original/baseline readmission probabilities without any interventions are the same across all care facilities. Specifically, we directly use the readmission probabilities for home patients to estimate the readmission probabilities for SNF and HS patients. Here, home patients' readmission risks are used as the reference to derive the conservative lower bounds of intervention effectiveness. That is, for SNF and HS patients, by assigning the lowest possible (lower bound) value for the baseline readmission risk, we are assuming patients are less likely to readmit without interventions. Hence, the effect of intervention is assumed to be at its smallest, thus yielding a conservative estimate of intervention effectiveness.

Given constant baseline readmission risk assumption, intervention is always more effective to the patients in lower level care facilities. That is, the following assumption holds for $t = 0, 1, ..., T-1$.

Assumption 6.7. Intervention is more effective to the patients in lower level of care facilities, i.e., the following inequalities should be satisfied for $i < j \in \{1, ..., S\}$:

$$p_t^W(0|i) - p_t^I(0|i) \leq p_t^W(0|j) - p_t^I(0|j),$$
$$\sum_{s'=k}^{S} p_t^I(s'|i) - \sum_{s'=k}^{S} p_t^W(s'|i) \leq \sum_{s'=k}^{S} p_t^I(s'|j) - \sum_{s'=k}^{S} p_t^W(s'|j).$$

Proposition 6.3 *Given Assumptions 6.1-6.5 and 6.7, there exists a monotone optimal policy $a^*(s_t)$ which is nondecreasing in s_t for $t = 0, 1, ..., T-1$. This implies there exists*

an $s_t^* \in \mathbb{S} \setminus \{0\}$ *such that*

$$a^*(s_t) = \begin{cases} W & \text{if } s_t \leq s_t^*, \\ I & \text{if } s_t > s_t^*. \end{cases} \tag{6.4}$$

Proof: See Appendix D. ∎

In the context of this problem, this monotone optimal policy can be interpreted as follows: for each decision epoch, the optimal action at the lowest level (least intense) care facility provides an upper bound for the optimal action at higher level care facilities. This implies that for a certain time epoch, if the optimal action for home patients is "Wait", the optimal actions for HS and SNF patients must be "Wait" as well. The proof of Proposition 3 is given in Theorem 4.7.4 in [101] and hence is omitted.

Scenario 2: Increasing baseline readmission risk in care facility level

In Scenario 2, we assume patients in higher level care facility have higher risk of getting readmitted even with necessary treatment provided at the care facilities. Thus, we are implicitly assuming riskier patients stay at a higher level care facility, and even with more intense level of treatment, their baseline readmission risk is higher. Supporting evidence can be referenced from medical literature similarly to Assumption 6.3. Here, we again use the home patients' readmission risk as the reference point. Specifically, we assume readmission risk increases by $\alpha = 0.02$ as care facility level increases by one. Hence, with $p_t^I(0|H)$ set as the reference, $p_t^I(0|HS)$ and $p_t^I(0|SNF)$ can be expressed as following: $p_t^I(0|HS) = p_t^I(0|H) + \alpha$, and $p_t^I(0|SNF) = p_t^I(0|H) + 2\alpha$.

For the transitions within care facilities, we assume the same proportion among the intervened transition probabilities is maintained for the baseline case. Hence, $p_t^I(1|s) : p_t^I(2|s) : p_t^I(3|s) = p_t^W(1|s) : p_t^W(2|s) : p_t^W(3|s)$.

6.4.3 Cost Parameter Estimation

Finally, we assign cost parameter values from analyzing the patient records and consulting the healthcare professionals at the collaborating hospital. For both scenarios, the following parameter values are assumed to obtain the optimal policies:

• $C_{CF}(s)$ is defined as the total care facility expenses accrued during a decision epoch, thus can be calculated by [average daily care facility cost $\times$ one decision epoch (=10 days)]. Specifically, the average daily care facility cost can be approximated by dividing the average total cost incurred during patients' entire stay at facility s by the average number of days patients stay at the facility. Here, the average total cost is estimated by consulting the healthcare professionals at the collaborating hospital, and the average duration of stay is calculated from analyzing the patient records.

• $V_{\text{Nonreadmit}}$ represents the value of a nonreadmitted patient at the end of the planning horizon and is approximated by the average readmission penalty per person.

6.5 Numerical Study

In this section, we present the implementation of the MDP model of Section 6.2 using the assumed parameter settings as shown in Section 6.4. The optimal policy for varying values of intervention cost C_I is presented in Figure 6.28 for both scenarios.

There are several interesting properties of the optimal policy. First, in accordance with intuition, as an intervention becomes more costly, the optimal region of "Wait" expands, thus it is more cost-effective to not give interventions. Second, as expected from Proposition 6.3, under Scenario 1, the optimal policy is a monotone policy and thus is a control-limit type (see Figure 6.28 (i) Scenario 1 where the optimal policy is divided

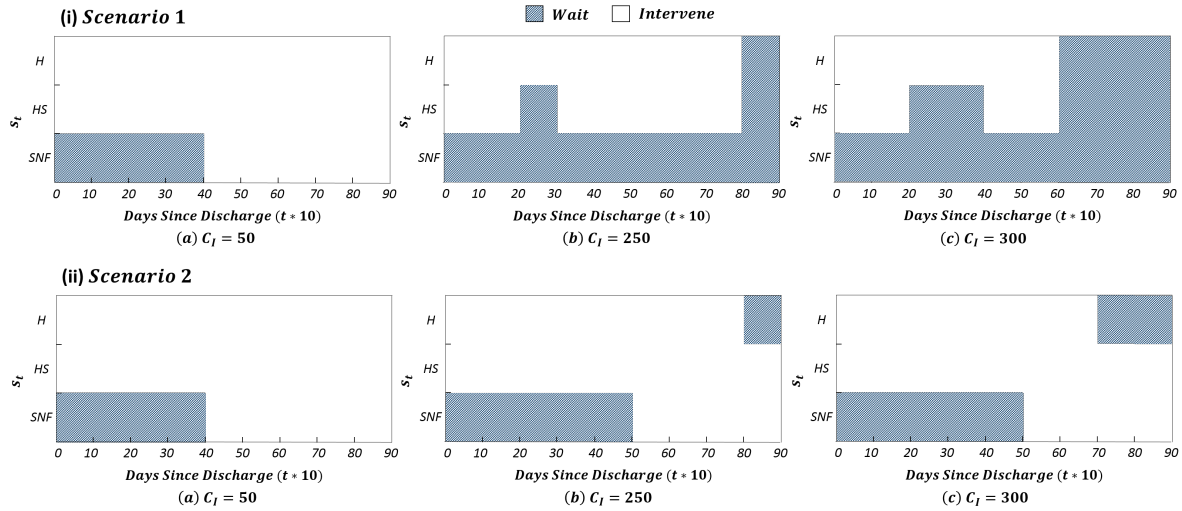

Figure 6.28: Illustrative examples of optimal policy for (i) Scenario 1 and (ii) Scenario 2 under varying intervention costs

into two regions). This implies that if the optimal action for patients at a care facility is "Wait" at time t, then the optimal action for higher level of care facilities at time t must be chosen as "Wait". However, for Scenario 2, this monotonicity no longer holds. As we can no longer assume intervention is more effective to either higher or lower level of care facility, the MDP no longer produces monotone policies. Thus, the varying effectiveness of interventions in different time epochs could explain the lack of monotonicity under Scenario 2. Finally, for Scenario 1, the optimal control-limit threshold s_t^* is not monotone with respect to time. This is a key insight from our model, which may in fact disagree with some common practices in many hospitals. In the medical community, intervention programs have been mainly focused on the immediate postoperative period [26] where prevailing literatures have shown the effectiveness of early rehabilitative programs on patient outcomes [88]. Most existing studies investigate the impact of PT during the immediate postsurgical period, i.e., examine the relationship between utilization of PT

services during the patient's inpatient stay (before discharge) and outcomes of care [88]. Hence, the effects of interventions conducted on the more far end of the recovery progress have not received much attention and are largely unknown. However, the optimal policy of our model suggests when intervention cost is high, interventions may be more cost-effective in the later days of recovery, i.e., see Figure 6.28 (i) Scenario 1(a) where the optimal policy suggests to start giving interventions to SNF patients after 40 days since discharge. This finding is clinically significant as it raises concerns in current common belief that interventions should be focused in early postacute phase of recovery.

6.5.1 Intervention Timing Effects

Our findings on the optimal intervention timing warrant further discussion. One possible explanation for such phenomena may be found in the causes of readmission where supporting evidence can be found in medical literatures. The causes of TJR readmission can be largely divided into two sources; surgical complications and medical complications [109]. Surgical complications include undesirable and unexpected result that occurs as a direct result of the surgery (e.g., surgical site infection (SSI), dislocation, noninfected draining wound, hematoma, pain management, deep vein thrombosis, etc.) while medical complications are not directly resulting from the surgery such as pneumonia, acute renal failure, nausea/vomiting, medication compliance, etc. Among various complications, CMS projects that only 20% of readmissions are preventable [25] where dislocation, SSI and postoperative hematoma are in many cases considered as the main preventable causes of readmission. Paper [109] identifies that while the overall readmission (all-cause readmission) mostly occurs in the early discharge phase, occurrence of the preventable causes of readmission is more shifted to the later phase of recovery. The types of interventions conducted at the collaborating hospital may be targeted (or more effective)

towards these preventable readmission causes, hence reducing readmission risk more effectively in the later phase of recovery. Moreover, it is asserted by [106] that patients discharged to SNF have higher rates of readmissions resulting from medical complications. Since major preventable readmission causes belong to surgical cases, SNF patients may be associated with more unpreventable readmission causes. This could explain why among the three care facilities, interventions are particularly not effective to the SNF patients in the early phase. Therefore, to better target SNF patients, the hospital needs to focus on devising specialized intervention programs targeting those early readmission causes. However, as early readmissions may not be preventable by postsurgery interventions only, attention should be also focused on the pre-surgery or surgical phase: identify high risk patients preoperatively and devise strategies to improve their modifiable risk factors (obesity, malnutrition, cardiovascular disease, smoking, etc.) or improve the surgical process quality.

In addition, as rehabilitation after joint replacement is dependent on the will and cooperation of the patients [110], patient adherence may play a key role in intervention effectiveness as well. Typically, patients' adherence to intervention programs at earlier time periods after surgery may not be strong due to pain, tiredness, anxiety, etc. Hence, appropriate peri-operative efforts should be supplemented to increase adherence and improve intervention outcomes, especially during the immediate postoperative period [111].

To further investigate the impact of optimal intervention timing, we compare the total cost resulting from our optimal policy to several practical guidelines that may be followed in practice. Specifically, we introduce 3 types of potential intervention policies for comparison:

• P1. Always Wait: Provide no interventions.

• P2. Always Intervene: Always provide interventions.

• P3. Early Interventions: Intervene only during the immediate postoperative period.

For P3, we further divide into varying durations of intervention:

- P3-1. Intervene up to first 10 days of discharge.
- P3-2. Intervene up to first 20 days of discharge.
- P3-3. Intervene up to first 30 days of discharge.

Note that P1 and P2 are more extreme types of policies and are mainly included for comparison purpose. P3-1, P3-2, and P3-3 represent more realistic strategies that may be conducted in practice.

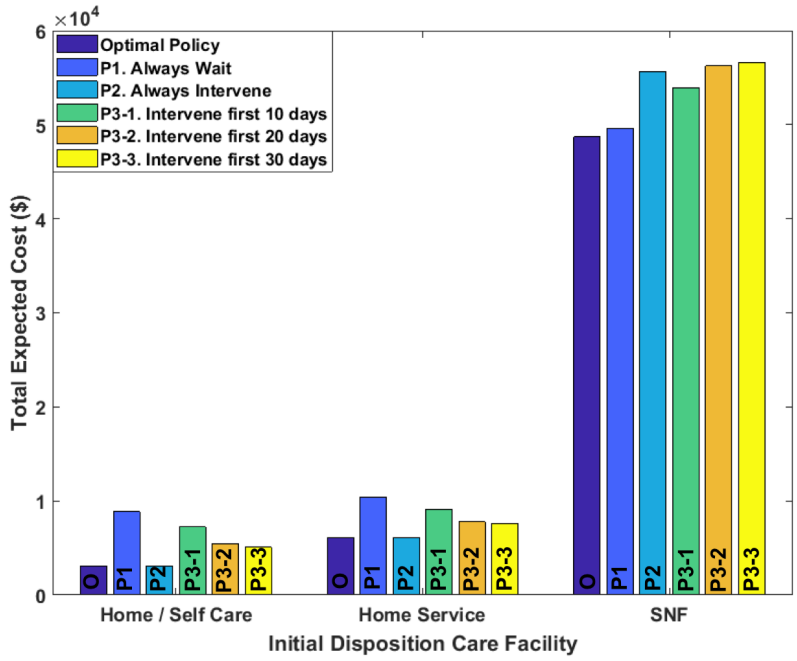

Figure 6.29: Total expected cost of varying intervention policies

The total expected cost for our optimal policy under Scenario 1 compared to the 5 intervention strategies is illustrated in Figure 6.29. An interesting observation is for SNF patients, providing no intervention yields lower cost than all other non-optimal

intervention policies; even the early intervention strategies yield significantly higher cost. This implies current intervention practice is not cost-effective for SNF patients, especially for the immediate postacute recovery period. This can also be explained through the causes of readmissions as SNF patients may have more unpreventable readmission causes. On the other hand, for home and home service patients, no intervention results in the worst case as the optimal policy is to provide interventions at all time epochs (see Figure 6.28 (i) Scenario 1(a)).

Comparing the early intervention strategies, the impact of intervention duration is more significant in the early phase. As intervention duration increases from first 10 days (P3-1) to first 20 days (P3-2), the total cost is reduced by 24% for home patients and 15% for home service patients. Cost reduction as duration increases from first 20 days (P3-2) to 30 days (P3-3) is 7% and 2% for home and home service respectively. Hence, interventions are more cost-effective in the early phase for both home and home service patients. SNF patients are more robust to intervention duration; the difference in cost is 4% and 0.6% for P3-1 to P3-2 and P3-2 to P3-3 respectively. Note that due to space limit, only the results of Scenario 1 are presented. However, since both Scenarios 1 and 2 yield similar result, the above analysis results can be extended to Scenario 2.

6.5.2 Intervention Effectiveness on Readmissions

Our optimal policy suggests to start giving interventions to SNF patients after the postacute recovery period. This can be mainly explained due to the varying effectiveness of interventions in different time periods; i.e., current intervention process at our collaborating hospital may not be highly effective to SNF patients in the immediate postoperative period. Hence, it can be expected that the intervention effectiveness plays a key role in determining the optimal policy. Therefore, the robustness of our model to intervention

effectiveness, i.e., how effective is the intervention in reducing patient's readmission risk, needs further investigation. There are two ways to carry out the experiment; one is to change the post-intervention readmission probability and the other is to change the pre-intervention (baseline) readmission risk. For our model, post-intervention readmission probabilities are mainly evaluated from the EHR data, whereas the baseline readmission risks rely more on assumptions. Specifically, the baseline readmission risks for home service and SNF patients are referenced from home patients' readmission probabilities. In Scenario 2 of Section 6.4, we assume the baseline readmission risk increases as care facility level increases, and the readmission risk difference among care facilities (α) is arbitrarily chosen as 0.02. Thus, the robustness of our model with respect to parameter α needs further investigation. Specifically, we experiment with values of α ranging from 0.01 to 0.04 where larger α implies interventions being more effective in reducing readmissions for higher level care facilities.

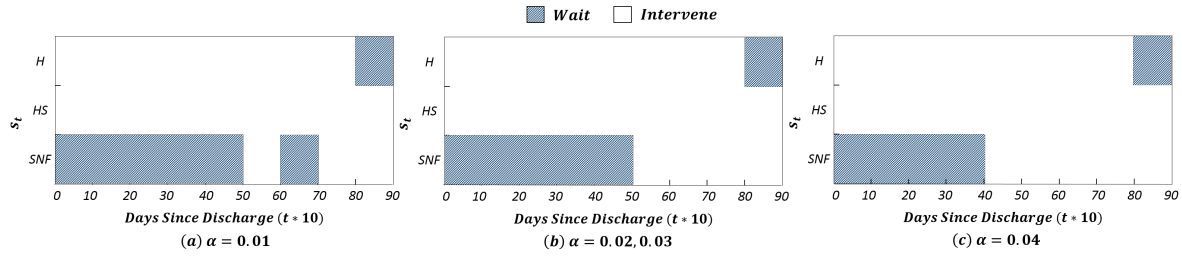

Figure 6.30: Sensitivity analysis of optimal policy with respect to baseline readmission risk difference α

The optimal policies for each α setting is presented in Figure 6.30. For lower level care facilities, the optimal policy is robust to α, i.e., the resulting optimal policies remain constant for home service patients. On the other hand, high level care facility is less robust to α. This is mainly because the higher the level of care facility, the more

impactful α value is, i.e., readmission risk difference is α between home service and home patients while $2 \times \alpha$ between SNF and home patients. In addition, we can observe from Figure 6.30 that as α increases, interventions become more effective, thus the "Intervene" region increases for SNF patients. Thus, more effective interventions yield optimal policies in which the optimal action is to intervene. This suggests an intuitive yet clinically meaningful insight to intervention protocols. Our optimal policy suggests current intervention process at our collaborating hospital may not be highly effective to SNF patients in the immediate postoperative period. However, as illustrated in Figure 6.30, as interventions become more effective (α increases), intervention starting day for SNF patients shifts closer to POD 0. Thus, efforts should be focused on devising interventions that are effective during the immediate postoperative period, especially for SNF patients.

6.5.3 Readmission Penalty Effects

Among the cost parameters used in our model, care facility cost and intervention cost are safe to assume to be constant throughout time, i.e., costs are not expected to change significantly over the years. However, readmission penalty can differ considerably each year as it is a function of both the hospital's current readmission rate and the national average performance, both of which are for a specific time period and are highly subject to vary. As the impact of readmission penalty can be significant on the total expected cost, sensitivity analysis is essential to check the robustness of our result. Specifically, we vary the readmission penalty by $0.5 \times R$ to assume both smaller and larger penalty settings. The resulting optimal policy when intervention cost C_I is assumed as 250 is presented in Figure 6.31.

As readmission penalty per person (R) increases, the optimal region of "Intervene"

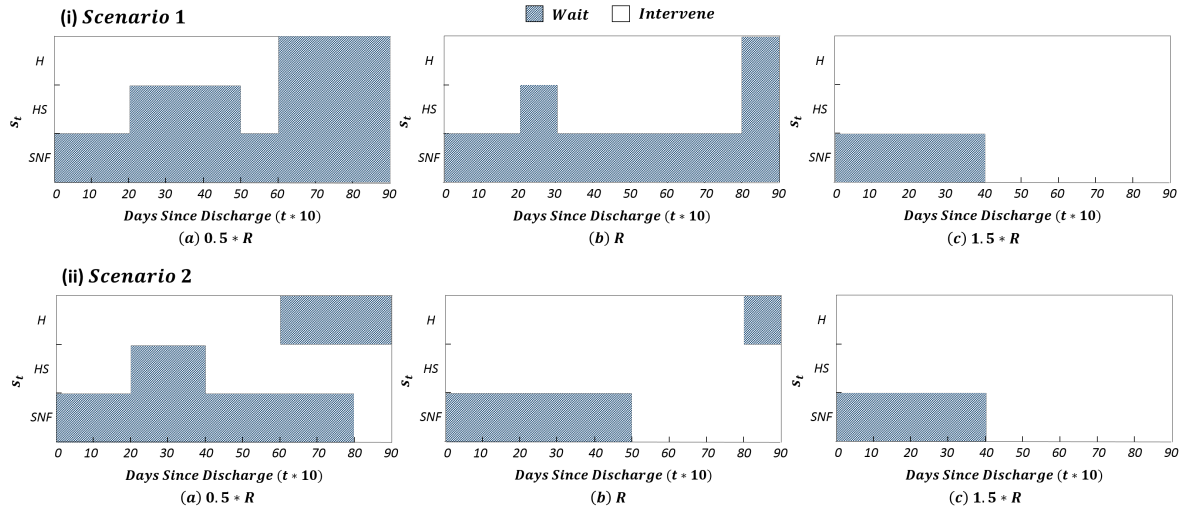

Figure 6.31: Sensitivity analysis of optimal policy with respect to readmission penalty

expands in accordance with intuition. This is because the benefit of preventing readmission and thus paying less penalty outweighs the cost of providing the interventions. We can observe the same trend for all care facility types under both scenarios. Another interesting observation is even when readmission penalty is 1.5 times larger than current cost, it is still optimal to not give interventions to SNF patients in the early postoperative period. However, we can observe that the intervention starting time shifts to POD 0 as larger readmission penalty is imposed, thus optimal policy follows a more conservative approach (i.e., give more interventions to avoid high penalty).

6.6 Conclusions

In this chapter, we formulate the optimal TJR postoperative intervention process as a finite-horizon discrete-time MDP, and investigate the structural properties to gain

insights of the optimal policy. A key insight we can obtain is that from the cost-effectiveness perspective, the common practice of focusing interventions on immediate postoperative periods needs further justification. Based on the data obtained from the collaborating hospital, the optimal policy suggests to start intervening SNF patients after the postacute recovery period. Although it may seem counterintuitive, possible explanation is that current interventions conducted at the hospital may not be so effective in reducing the early readmissions. Thus, more attention should be focused on surgical or pre-surgical interventions such as improving the surgical procedure quality or identifying high risk patients before surgery to improve their modifiable risk factors. Potentially, our research could serve as a guide for optimal postdischarge decisions where intervention resources or budget are limited. With the modeling framework, hospital policymakers could estimate the total expenses related to joint replacement postdischarge services, or assess the cost-effectiveness of varying intervention policies.

Chapter 7

Conclusions and Extensions

7.1 Summary of Contributions

This disseration introduces various analytical tools and models to improve the efficiency of the healthcare system across the care delivery cycle. Specifically, we present a set of techniques and strategies that can be used by healthcare professionals to reduce the inefficiencies or delays at each stage of the care delivery cycle: prevention, diagnosis, treatment, home care. Such tools would ideally improve health system by: (1) improving access to care; (2) reducing unnecessary waiting time or blocking; and (3) reducing cost by optimizing resource allocation. To conclude this dissertation, we discuss how the models or tools introduced in each chapter contributes to such operational improvement.

Improving access to care: Chapter 3 introduces a Markov chain based modeling framework to assess the impact of implementing a new service model in primary care clinics. Through an application study at a community hospital, we show that the implementation of joint visit service model leads to a significant improvement in patient throughput which can help improve patient access substantially. Moreover, we compare the impact of different workload allocation policies on system performance, and provide managerial insights on optimally redistributing workload among care providers.

Reducing unnecessary waiting time or blocking: Chapters 4 and 5 investigate systematic and evidence-based decisions that can lead to less delayed or congested system. In Chapter 4, a system-theoretic method is introduced to analyze the diagnosis-to-treatment process for lung cancer patients. As the lung cancer detection, diagnosis, and selection of the most appropriate treatment can be difficult, it is common to observe frequent and potentially harmful delays. Therefore, we pursue a rigorously quantitative approach to identify the most impeding wait times (bottlenecks) along the care delivery process to facilitate quality improvement by directing attention to specific opportuities. Such a method provides lung cancer specialists and caregivers a quantitative tool to study and reduce unnecessary wait times. While Chapter 4 focuses on the sources of delay during the diagnosis phase, Chapter 5 directs attention to the delays or blocking for hospitalized patients (i.e., treatment phase). Specifically, we investigate the patient flow in hospitals since improving it can have a significant impact on quality of care as well as patient satisfaction. To do so, a finite capacity queueing network model based iterative procedure is formulated to evaluate the complex interactions among multiple units within a hospital. Utilizing such a framework, we investigate system properties to provide managerial guidance on improving patient transitions by focusing on bed capacity allocation, external admission scheduling, and service time variabilty control.

Reducing cost by optimizing resource allocation: Chapter 6 formulates the TJR post-operative intervention process as a finite-horizon discrete-time MDP to determine the optimal timing and target group of interventions. Utilizing such a framework, health professionals can optimize postoperative patient care and better allocate hospital intervention resources with minimal expense, i.e., to minimize the total cost, when should the hospital provide interventions to which group of patients. Optimal policies derived from such models can serve as a clinical practice guideline in developing appropriate

postoperative care process.

7.2 Extensions

This section describes two potential extensions to the work introduced in this disseration. The first problem considers the transient behavior of transition delay times to help hospitals better cope with disastrous events, and the second problem relates to integrating risk predictive modeling to provide a more personalized decision support tool. The details of each problem are described in the following subsections.

7.2.1 Extension 1: Analysis of Transient Behavior

In Chapter 5, the mean and variability of the inpatient transition queue times are evaluated using steady state analysis. This is a reasonable approach when long-term patient flows are considered such as in normal hospital operations setting. However, if we were to estimate the hospital capacity during or after a disastrous event, transient analysis would be more capable of representing real-time capacity estimation and patient waiting times. When a disaster occurs, the ED admissions are expected to increase 3-5 times the normal volume, which could easily overwhelm the hospital's resources [112]. In such disaster situations, emergency preparedness, which can be achieved through real-time hospital capacity estimation, helps hospitals to cope with sudden surge of patients by temporarily reallocating resources among different units. As this sudden surge of patients typically lasts only for a short period of time without reaching a steady-state, long-term performance evaluation of normal operations becomes inadequate in disaster modeling. Hence, understanding the transient behavior of hospital capacity and patient wait times is vital to provide timely treatment and improve safety to both the patients

injured in disasters and the regular inpatients.

7.2.2 Extension 2: Integration of Risk Predictive Modeling

Another worthwhile area for further research is the extension of the TJR-postoperative intervention process model towards a more personalized patient-specific decision model. In Chapter 6, readmission risks differ depending on where the patient is located at (care facility type) and how many days have elapsed since discharge. This is largely based on the assumption that care facility level typically matches with the patient's risk level; higher risk patients stay at higher level of care facility. In general, this holds true, which allows to perform cost-effectiveness analysis on an average population level. However, if we were to provide a more patient-specific decision support tool, a possible approach is to integrate risk prediction into the MDP modeling framework. This can be achieved by first developing a readmission risk predictive model to stratify patients into different risk groups, and then investigating the optimal intervention policy for each risk group. The idea of integrating readmission risk prediction model and intervention process model is presented in [113] where we stratify patients into different risk groups according to the predictive model results, and analyze the impact of the initial disposition care facility type on each risk group. By integrating risk predictive modeling to the current MDP framework, we can move one step closer to providing personalized optimal intervention plans.

Appendices

Appendix A. Proofs of Chapter 3

First, the state transitions and the corresponding balance equations are derived for each system in the following order: Current System, Joint Visit System with Provider Wrap-up and Two MAs, Joint Visit with MA Wrap-up and Two MAs, Joint Visit System with Provider Wrap-up and Three MAs, and Joint Visit System with MA Wrap-up and Three MAs.

Current System: The state transitions are illustrated in Figure A.1. The system

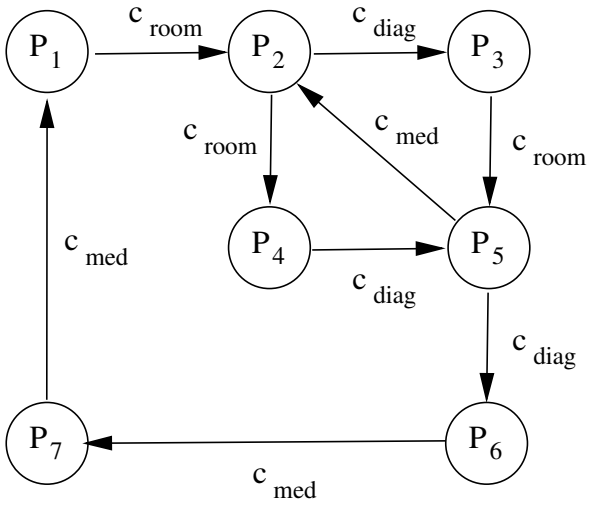

Figure A.1: Transition diagram of current system

balance equations are:

$$
\begin{aligned}
c_{\text{room}} P_1 &= c_{\text{med}} P_7, \\
(c_{\text{room}} + c_{\text{diag}}) P_2 &= c_{\text{room}} P_1 + c_{\text{med}} P_5, \\
c_{\text{room}} P_3 &= c_{\text{diag}} P_2, \\
c_{\text{diag}} P_4 &= c_{\text{room}} P_2, \\
(c_{\text{diag}} + c_{\text{med}}) P_5 &= c_{\text{room}} P_3 + c_{\text{diag}} P_4, \\
c_{\text{med}} P_6 &= c_{\text{diag}} P_5, \\
c_{\text{med}} P_7 &= c_{\text{med}} P_6.
\end{aligned}
\tag{A.1}
$$

In addition,

$$
\sum_{i=1}^{7} P_i = 1. \tag{A.2}
$$

Solving the above equations, we obtain the steady state probabilities P_i, $i = 1, \ldots, 7$.

Joint Visit System with Provider Wrap-up and Two MAs: The states are defined as $S = (s_{\text{room}},\ b_{\text{joint}},\ s_{\text{joint}},\ s_{\text{wrap}},\ b_{\text{med}},\ s_{\text{med}})$. Note that there is no state b_{wrap} since the provider conducts wrap-up immediately after the joint visit. This results in a total of 6 states defined as following:

$$
S_1 = (1,0,0,1,0,0), \quad S_2 = (1,0,0,0,1,0), \quad S_3 = (0,0,0,0,1,1),
$$

$$
S_4 = (0,1,0,1,0,0), \quad S_5 = (0,0,1,0,1,0), S_6 = (0,0,0,1,0,1).
$$

The state transitions are illustrated in Figure A.2. The following balance equations can

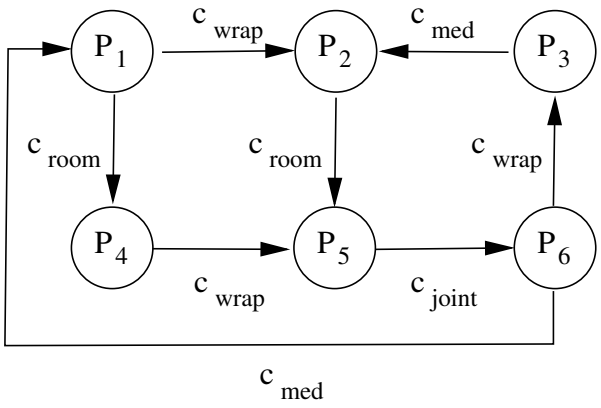

Figure A.2: Transition diagram of joint system with two MAs and provider wrap-up

be derived:

$$\begin{aligned}
(c_{\text{room}} + c_{\text{wrap}})P_1 &= c_{\text{med}}P_6, \\
c_{\text{room}}P_2 &= c_{\text{wrap}}P_1 + c_{\text{med}}P_3, \\
c_{\text{med}}P_3 &= c_{\text{wrap}}P_6, \\
c_{\text{wrap}}P_4 &= c_{\text{room}}P_1, \\
c_{\text{joint}}P_5 &= c_{\text{room}}P_2 + c_{\text{wrap}}P_4, \\
(c_{\text{wrap}} + c_{\text{med}})P_6 &= c_{\text{joint}}P_5, \\
\sum_{i=1}^{6} P_i &= 1.
\end{aligned} \tag{A.3}$$

Solving these equations we obtain the steady state probabilities P_i, $i = 1, \ldots, 6$. The performance of the system can be evaluated through these probabilities.

Joint Visit System with MA Wrap-up and Two MAs: Define states as $S = (s_{\text{room}}, s_{\text{joint}}, s_{\text{med}})$. This results in a total of 3 states defined as following:

$$S_1 = (1, 0, 0, 0), \quad S_2 = (0, 0, 1, 0), \quad S_3 = (0, 0, 0, 1).$$

The state transitions are illustrated in Figure A.3 and the balance equations are:

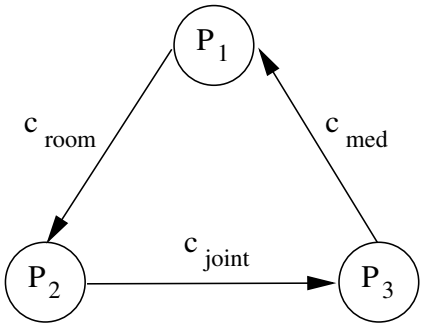

Figure A.3: Transition diagram of joint system with two MAs and MA wrap-up

$$
\begin{aligned}
c_{\text{room}} P_1 &= c_{\text{med}} P_3, \\
c_{\text{joint}} P_2 &= c_{\text{room}} P_1, \\
c_{\text{med}} P_3 &= c_{\text{joint}} P_2, \\
\sum_{i=1}^{3} P_i &= 1.
\end{aligned}
\tag{A.4}
$$

Solving these equations we obtain the steady state probabilities P_i, $i = 1, 2, 3$. The performance of the system can be evaluated through these probabilities.

Joint Visit System with Provider Wrap-up and Three MAs: Assume there are two MAs supporting each provider. The state transitions are illustrated in Figure A.4. In addition, the following balance equations can be derived as:

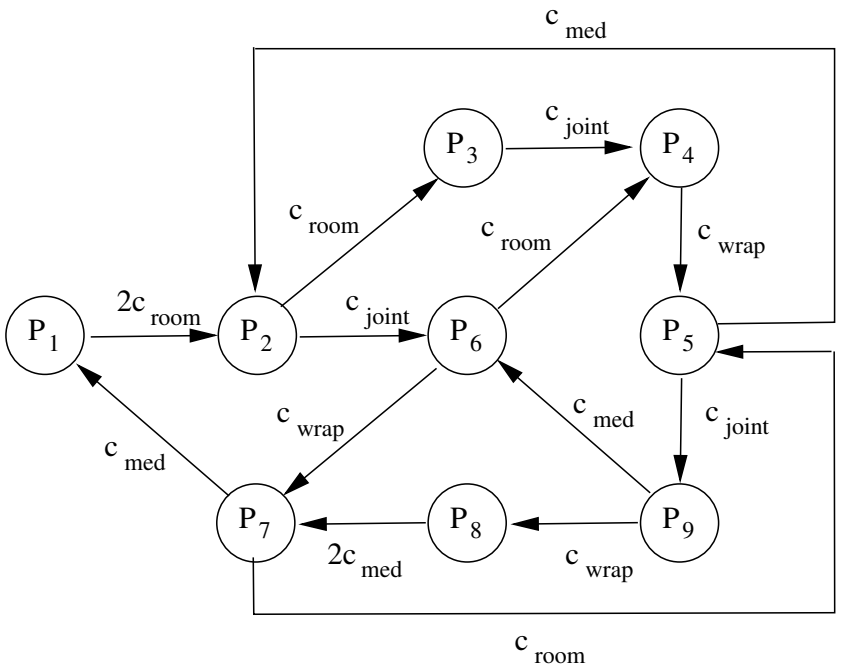

Figure A.4: Transition diagram of joint visit system with Provider wrap-up

$$
\begin{aligned}
2c_{\text{room}}P_1 &= c_{\text{med}}P_7,\\
(c_{\text{room}} + c_{\text{joint}})P_2 &= 2c_{\text{room}}P_1 + c_{\text{med}}P_5,\\
c_{\text{joint}}P_3 &= c_{\text{room}}P_2,\\
c_{\text{wrap}}P_4 &= c_{\text{joint}}P_3 + c_{\text{room}}P_6,\\
(c_{\text{joint}} + c_{\text{med}})P_5 &= c_{\text{wrap}}P_4 + c_{\text{room}}P_7,\\
(c_{\text{room}} + c_{\text{wrap}})P_6 &= c_{\text{joint}}P_2 + c_{\text{med}}P_9,\\
(c_{\text{room}} + c_{\text{med}})P_7 &= c_{\text{wrap}}P_6 + 2c_{\text{med}}P_8,\\
2c_{\text{med}}P_8 &= c_{\text{wrap}}P_9,\\
(c_{\text{wrap}} + c_{\text{med}})P_9 &= c_{\text{joint}}P_5,\\
\sum_{i=1}^{9} P_i &= 1.
\end{aligned}
\tag{A.5}
$$

Solving these equations we obtain the steady state probabilities P_i, $i = 1, \ldots, 9$.

Joint Visit System with MA Wrap-up and Three MAs: Assume there are two MAs supporting each provider. Figure A.5 illustrates the state transitions, and the balance equations are derived as follows:

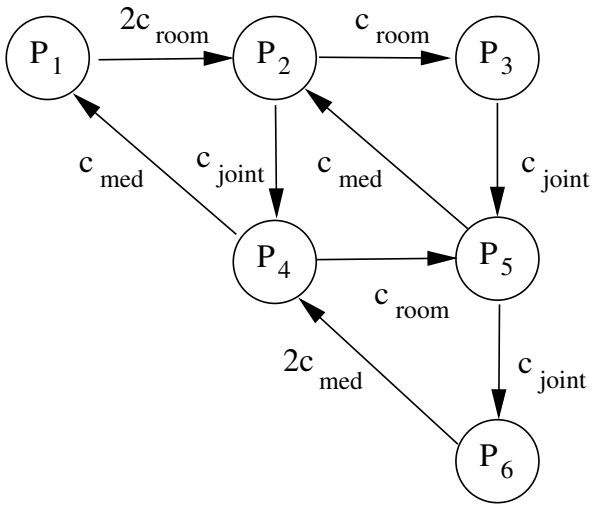

Figure A.5: Transition diagram of joint visit system with MA wrap-up

$$
\begin{aligned}
2c_{\text{room}}P_1 &= c_{\text{med}}P_4, \\
(c_{\text{room}} + c_{\text{joint}})P_2 &= 2c_{\text{room}}P_1 + c_{\text{med}}P_5, \\
c_{\text{joint}}P_3 &= c_{\text{room}}P_2, \\
2c_{\text{med}}P_6 &= c_{\text{joint}}P_4, \\
(c_{\text{room}} + c_{\text{med}})P_4 &= c_{\text{joint}}P_2 + 2c_{\text{med}}P_6, \\
(c_{\text{joint}} + c_{\text{med}})P_5 &= c_{\text{joint}}P_3 + c_{\text{room}}P_4, \\
2c_{\text{med}}P_6 &= c_{\text{joint}}P_5, \\
\sum_{i=1}^{6} P_i &= 1.
\end{aligned}
\tag{A.6}
$$

The proof of Proposition 3.2 requires the following lemmas.

Lemma 7.1 *In the joint visit system with three MAs and provider wrap-up, the probability that both MAs are busy in a provider team is monotonically decreasing with respect to the service rate of MA visit, i.e.,*

$$
\frac{\partial \alpha^i}{\partial c^i_{\text{med}}} \leq 0, \quad i = \text{MD}, \text{PA}. \tag{A.7}
$$

Lemma 7.2 *In Procedure 3.1, if* $\widetilde{c}^{i,n}_{\text{med}} < \widetilde{c}^{i,n-1}_{\text{med}}$, $i =$*MD, PA, then* $\alpha^{i,n} \geq \alpha^{i,n-1}$.

Proof of Lemma 7.1: The proof of this lemma is based on the results obtained from *Mathematica*. From balance equations (A.5), the steady state probabilities P_i, $i = 1, \ldots, 9$, can be obtained. Then probability α and partial derivative $\partial\alpha/\partial c_{\text{med}}$ can be derived. Under the constraints $\forall c_i > 0$, the statement of $\partial\alpha/\partial c_{\text{med}} \leq 0$ is verified as "TRUE" by *Mathematica*. ■

Proof of Lemma 7.2: Induction is used in this proof. When $n = 1$, from (3.8),

$$\widetilde{c}_{\text{med}}^{\text{MD},1} = c_{\text{med}}^{\text{MD},0}(1 - \alpha^{\text{PA},0}/2),$$

we obtain $\widetilde{c}_{\text{med}}^{\text{MD},1} \leq c_{\text{med}}^{\text{MD},0}$. Then by Lemma 7.1, we have

$$\alpha^{\text{MD},1} \geq \alpha^{\text{MD},0}.$$

Similarly, from (3.9),

$$\widetilde{c}_{\text{med}}^{\text{PA},1} = c_{\text{med}}^{\text{PA},0}(1 - \alpha^{\text{MD},1}/2),$$

we obtain $\widetilde{c}_{\text{med}}^{\text{PA},1} \leq c_{\text{med}}^{\text{PA},0}$ and by Lemma 7.1,

$$\alpha^{\text{PA},1} \geq \alpha^{\text{PA},0}.$$

The base case is proved. Now assume when $n = l$,

$$\widetilde{c}_{\text{med}}^{i,l} < \widetilde{c}_{\text{med}}^{i,l-1}, i = \text{MD}, \text{PA},$$

$$\alpha^{i,l} \geq \alpha^{i,l-1}.$$

Next when $n = l + 1$, from (3.8) and Lemma 7.1, we obtain

$$\widetilde{c}_{\text{med}}^{\text{MD},l+1} = c_{\text{med}}^{\text{MD},0}(1 - \alpha^{\text{PA},l}/2) \leq c_{\text{med}}^{\text{MD},0}(1 - \alpha^{\text{PA},l-1}/2) = \widetilde{c}_{\text{med}}^{\text{MD},l}.$$

It is clear that if $\widetilde{c}_{\text{med}}^{\text{MD},l+1} = \widetilde{c}_{\text{med}}^{\text{MD},l}$, then $\widetilde{P}_j^{\text{MD},l+1} = \widetilde{P}_j^{\text{MD},l}$, $j = 1, \ldots, 9$. Thus

$$\alpha^{\text{MD},l+1} = \alpha^{\text{MD},l}.$$

If $\widetilde{c}_{\text{med}}^{\text{MD},l+1} \leq \widetilde{c}_{\text{med}}^{\text{MD},l}$, by Lemma 7.1,

$$\alpha^{\text{MD},l+1} \geq \alpha^{\text{MD},l}.$$

Therefore, in both cases, we obtain $\alpha^{\text{MD},l+1} \geq \alpha^{\text{MD},l}$. Next, from (3.9) and Lemma 7.1, we obtain

$$\widetilde{c}_{\text{med}}^{\text{PA},l+1} = c_{\text{med}}^{\text{PA},0}(1 - \alpha^{\text{MD},l+1}/2) \leq c_{\text{med}}^{\text{PA},0}(1 - \alpha^{\text{PA},l}/2) = \widetilde{c}_{\text{med}}^{\text{PA},l}.$$

Again if $\widetilde{c}_{\text{med}}^{\text{PA},l+1} = \widetilde{c}_{\text{med}}^{\text{PA},l}$, then $\widetilde{P}_j^{\text{PA},l+1} = \widetilde{P}_j^{\text{PA},l}$, $j = 1, \ldots, 9$. Thus

$$\alpha^{\text{PA},l+1} = \alpha^{\text{PA},l}.$$

If $\widetilde{c}_{\text{med}}^{\text{PA},l+1} \leq \widetilde{c}_{\text{med}}^{\text{PA},l}$, by Lemma 7.1,

$$\alpha^{\text{PA},l+1} \geq \alpha^{\text{PA},l}.$$

Again in either case, $\alpha^{\text{PA},l+1} \geq \alpha^{\text{PA},l}$. By induction, the argument follows. ■

Proof of Proposition 3.2: From Lemma 7.2, $\alpha^{i,n}$ is monotonically increasing. Since it is bounded between 0 and 1, it is convergent. From (3.8)-(3.10), $\widetilde{P}_j^{i,n}$ reaches its limit when $n \to \infty$, and so does $TP^{i,n}$. Thus, Procedure 3.1 is convergent. ■

Proof of Proposition 3.3: When a new patient is not roomed right after a patient leaves, only one patient will exist in a provider's system. In this case, a waiting time occurs for the patient to get roomed. Such time is denoted as w and can be calculated by the decreased amount of MA's service rate.

$$w^i = \frac{1}{\widetilde{c}_{\text{med}}^i} - \frac{1}{c_{\text{med}}^i}, \quad i = \text{MD}, \text{PA}.$$

Assume the proportion of time when there is only one patient in the system is w out of average patient length of visit LOV. Denote this proportion as r, i.e.,

$$r = \frac{w}{LOV}. \tag{A.8}$$

Then, the average number of patients can be evaluated as

$$N = 1 \cdot r + 2 \cdot (1 - r) = 2 - r.$$

Using Little's Law, the average patient length of visit, LOV, can also be represented by

$$LOV = \frac{N}{TP} = \frac{2 - r}{TP}. \tag{A.9}$$

From equations (A.8) and (A.9), the ratio r can be formulated as a quadratic equation.

$$r^2 - 2r + w \cdot TP = 0. \tag{A.10}$$

It can be shown that, there always exists at least one real root for equation (A.10), which equals to

$$r = 1 - \sqrt{1 - TP \cdot w}, \tag{A.11}$$

where TP and w are obtained from (3.11) and (3.13), respectively.

Note that $TP \cdot w$ can be formulated using r as following:

$$TP \cdot w = \frac{2 - r}{LOV} \cdot r \cdot LOV = r(2 - r)$$

Now the discriminant of equation (A.10) can be represented by r,

$$Discriminant = 4(1 - TP \cdot w) = 4(1 - 2r + r^2) = 4(r - 1)^2 \geq 0$$

As $Discrminant \geq 0$ always, we can conclude equation (A.10) has at least one real root. Using quadratic formula gives:

$$r = 1 \pm \sqrt{1 - TP \cdot w}$$

But since r is a proportion and thus must be in the range of [0,1], we can conclude the feasible root to equation (A.10) is:

$$r = 1 - \sqrt{1 - TP \cdot w}.$$

Note that if $r = 1$, there exists only one solution, and this implies there is always one patient in the system. ■

Proof of Proposition 3.5: Using *Mathematica*, the following statements are true: $\forall i, j$,

$$\begin{aligned}
\frac{\partial TP^i}{\partial c^i_j} &\geq 0, \quad i = \text{MD}, \text{PA}, \quad j = \text{room}, \text{joint}, \text{wrap}, \text{med}, \\
\frac{\partial \rho_i}{\partial c^i_j} &\geq 0, \quad i = \text{MD}, \text{PA}, \quad j = \text{room}, \text{med}, \\
\frac{\partial \rho_i}{\partial c^i_j} &\leq 0, \quad i = \text{MD}, \text{PA}, \quad j = \text{joint}, \text{wrap}, \\
\frac{\partial \rho_{MA}}{\partial c^i_j} &\leq 0, \quad i = \text{MD}, \text{PA}, \quad j = \text{room}, \text{med}, \\
\frac{\partial \rho_{MA}}{\partial c^i_{\text{wrap}}} &\geq 0, \quad i = \text{MD}, \text{PA}.
\end{aligned}$$

The monotonicity arguments follow immediately. ∎

Proof of Proposition 3.6: Using *Mathematica*, the following statements are true: $\forall i, j$,

$$\begin{aligned}
\frac{\partial TP^i}{\partial c^i_j} &\geq 0, \quad i = \text{MD}, \text{PA}, \quad j = \text{room}, \text{joint}, \text{med}, \\
\frac{\partial \rho_i}{\partial c^i_j} &\geq 0, \quad i = \text{MD}, \text{PA}, \quad j = \text{room}, \text{med}, \\
\frac{\partial \rho_i}{\partial c^i_{\text{joint}}} &\leq 0, \quad i = \text{MD}, \text{PA}.
\end{aligned}$$

The monotonicity arguments follow immediately. ∎

Appendix B. Proofs of Chapter 4

Proof of Proposition 4.1: Due to the independence assumption, all the covariances become zero when a serial process is considered. Thus, for a serial process i, its variance

can be evaluated as

$$\begin{aligned} Var_i &= E\Big[\Big(\sum_{j=1}^{16}\theta_{i,j}\mathcal{W}_j\Big)^2\Big] - E^2\Big[\sum_{j=1}^{16}\theta_{i,j}\mathcal{W}_j\Big] \\ &= \sum_{j=1}^{16}\theta_{i,j}E[\mathcal{W}_j^2] - \Big(\sum_{j=1}^{16}\theta_{i,j}E[\mathcal{W}_j]\Big) \\ &= \sum_{j=1}^{16}\theta_{i,j}[Var(\mathcal{W}_j) + \mathcal{T}_j^2] - \Big(\sum_{j=1}^{16}\theta_{i,j}\mathcal{T}_j\Big) \\ &= \sum_{j=1}^{16}\theta_{i,j}[\mathcal{V}_j^2\mathcal{T}_j^2 + \mathcal{T}_j^2] - \Big(\sum_{j=1}^{16}\theta_{i,j}\mathcal{T}_j\Big). \end{aligned}$$

Considering the weighted average and normalization, we obtain

$$Var = \frac{\sum_{i=1}^{K} p_i Var_i}{\sum_{i=1}^{K} p_i}.$$

By taking $CV = \sqrt{Var}/T$, the argument follows. ∎

Proof of Proposition 4.2: First, consider partial derivative with respect to $\tau_{i,j}$, which corresponds to element $\mathcal{T}_k$.

$$\begin{aligned} \frac{\partial T}{\partial \tau_{i,j}} = \frac{\partial T}{\partial \mathcal{T}_k} &= \frac{\partial \frac{\sum_{i=1}^{K}\sum_{j=1}^{16}\theta_{i,j}\mathcal{T}_j}{\sum_{i=1}^{K} p_i}}{\partial \mathcal{T}_k} = \frac{\sum_{i=1}^{K} p_i\theta_{i,k}}{\sum_{i=1}^{K} p_i} \\ &> 0. \end{aligned} \tag{B.1}$$

Next, for partial derivative with respect to $cv_{i,j}$, it corresponds to element $\mathcal{V}_k$.

$$\begin{aligned}
\frac{\partial CV}{\partial cv_{i,j}} &= \frac{\partial CV}{\partial \mathcal{V}_k} = \frac{\partial \frac{\sqrt{\frac{\sum_{i=1}^{K} p_i Var_i}{\sum_{i=1}^{K} p_i}}}{T}}{\partial \mathcal{V}_k} \\
&= \frac{\partial \frac{\sqrt{\frac{\sum_{i=1}^{K} p_i \Big[\sum_{j=1}^{16} \theta_{i,j} \mathcal{T}_j^2 (\mathcal{V}_j^2+1) - \Big(\sum_{j=1}^{16} \theta_{i,j} \mathcal{T}_j\Big)^2\Big]}{\sum_{i=1}^{K} p_i}}}{T}}{\partial \mathcal{V}_k} \\
&= \frac{1}{\frac{\sqrt{\sum_{i=1}^{K} p_i}}{T}} \cdot \frac{1}{2} \cdot \Big(2 \sum_{i=1}^{K} \theta_{i,k} \mathcal{T}_k^2 \mathcal{V}_k\Big) \\
&\quad \Big/ \Big(\sum_{i=1}^{K} p_i \Big[\sum_{j=1}^{16} \theta_{i,j} \mathcal{T}_j^2 (\mathcal{V}_j^2+1) - \Big(\sum_{j=1}^{16} \theta_{i,j} \mathcal{T}_j\Big)^2 \Big] \Big)^{\frac{1}{2}} \\
&= \frac{1}{T \cdot CV \cdot \sum_{i=1}^{K} p_i} \cdot \Big(\sum_{i=1}^{K} p_i \theta_{i,k} \mathcal{T}_k^2 \mathcal{V}_k \Big) \qquad \text{(B.2)} \\
&> 0.
\end{aligned}$$

■

Proof of Proposition 4.3: Consider

$$\begin{aligned}
&T(\mathcal{T}_k - \delta_\tau \mathcal{T}_k) \\
&\quad = \frac{\sum_{i=1}^{K} p_i \Big[\sum_{j=1, j\neq k}^{M} \theta_{i,j} \mathcal{T}_j + \theta_{i,k} (\mathcal{T}_k - \delta_\tau \mathcal{T}_k) \Big]}{\sum_{i=1}^{K} p_i} \\
&\quad = \frac{\sum_{i=1}^{K} p_i \Big[\sum_{j=1}^{M} \theta_{i,j} \mathcal{T}_j - \delta_\tau \theta_{i,k} \mathcal{T}_k \Big]}{\sum_{i=1}^{K} p_i} \\
&\quad = T - \frac{\sum_{i=1}^{K} p_i \delta_\tau \theta_{i,k} \mathcal{T}_k}{\sum_{i=1}^{K} p_i},
\end{aligned}$$

Since T, δ_τ, and denominator $\sum_{i=1}^{K} p_i$ are common terms, the dominating factor is the enumerator, $\sum_{i=1}^{K} p_i \theta_{i,k} \mathcal{T}_k$. Reversing the sign of the term, the argument follows. ■

Proof of Proposition 4.4: From (B.2), we obtain

$$\frac{\partial CV}{\partial \mathcal{V}_k} = \frac{1}{T \cdot CV \cdot \sum_{i=1}^{K} p_i} \cdot \left(\sum_{i=1}^{K} p_i \theta_{i,k} \mathcal{T}_k^2 \mathcal{V}_k \right),$$

The denominator is a common term, thus the expression $\sum_{i=1}^{K} p_i \theta_{i,k} \mathcal{T}_k^2 \mathcal{V}_k$ is the dominating one. The argument follows immediately. ∎

The proof of Proposition 4.5 requires Lemma 7.3.

Lemma 7.3 *If $\{X_i, i = 1, \ldots, n\}$ are independently distributed gamma random variables with mean μ_i and standard deviation σ_i, then*

$$G(\omega_d) = \prod_{i=1}^{n} \left(\frac{\beta_{min}}{\beta_i} \right)^{\alpha_i} \sum_{k=0}^{\infty} \frac{\delta_k \gamma(\rho + k, \omega_d / \beta_{min})}{\Gamma(\rho + k)}, \tag{B.3}$$

where β_{min}, β_i, α_i, δ_k, ρ, $\gamma(\cdot)$, and $\Gamma(\cdot)$ are defined in (4.11).

Proof of Lemma 7.3: If $\{X_i, i = 1, \ldots, n\}$ are independently distributed gamma random variables, the density of $Y = X_1 + \ldots + X_n$ can be expressed as following (see [114]):

$$g(y) = \prod_{i=1}^{n} (\beta_{min}/\beta_i)^{\alpha_i} \sum_{k=0}^{\infty} \frac{\delta_k y^{\rho+k-1} e^{-y/\beta_{min}}}{\Gamma(\rho + k) \beta_1^{\rho+k}}, \quad y > 0, \tag{B.4}$$

where

$$\beta_{min} = \min(\beta_i), \quad \rho = \sum_{i=1}^{n} \alpha_i.$$

The coefficients δ_k can be obtained recursively by the formula:

$$\delta_{k+1} = \frac{1}{k+1} \sum_{i=1}^{k+1} i \nu_i \delta_{k+1-i}, \quad k = 1, 2, \ldots, \tag{B.5}$$

with

$$\delta_0 = 1,$$

$$\nu_k = \sum_{i=1}^{n} \alpha_i (1 - \beta_{min}/\beta_i)^k / k, \quad k = 1, 2, \ldots.$$

Then the CDF $G(w) = Pr(Y \leq \omega_d)$ can be obtained by integrating the probability density function $g(y)$.

$$\begin{aligned} G(\omega_d) &= \int_0^{\omega_d} g(y)dy \\ &= \prod_{i=1}^{n} \left(\frac{\beta_{min}}{\beta_i}\right)^{\alpha_i} \sum_{k=0}^{\infty} \frac{\delta_k \int_0^{\Omega_d} y^{\rho+k-1} e^{-y/\beta_{min}}}{\Gamma(\rho+k)\beta_1^{\rho+k}} dy. \end{aligned}$$

Using $a = c - 1$ and $y = bt$, then $\int_0^{\omega_d} y^a e^{-y/b} dy$ can be represented by the following:

$$\begin{aligned} \int_0^{\omega_d} y^a e^{-y/b} dy &= b^c \int_0^{w/b} t^{c-1} e^{-t} dt \\ &= b^c \cdot \gamma(c, w/b) \\ &= b^{a+1} \cdot \gamma(a+1, w/b), \end{aligned}$$

where $\gamma(a+1, \omega_d/b)$ is the lower incomplete gamma function, defined as

$$\gamma(a, x) = \int_0^x y^{a-1} e^{-y} dy,$$

and gamma function $\Gamma(a)$ can be related to the lower incomplete gamma function $\gamma(a, x)$ such that

$$\Gamma(a) = \lim_{x \to \infty} \gamma(a, x).$$

The formula for the integral term $\int_0^{\omega_d} y^{\rho+k-1} e^{-y/\beta_{min}} dy$ can be obtained.

$$\int_0^{\omega_d} y^{\rho+k-1} e^{-y/\beta_{min}} dy = \beta_{min}^{\rho+k} \cdot \gamma(\rho+k, w/\beta_{min}).$$

Substituting the integral term gives the following cumulative distribution function.

$$\begin{aligned} G(\omega_d) &= \prod_{i=1}^{n} \left(\frac{\beta_{min}}{\beta_i}\right)^{\alpha_i} \sum_{k=0}^{\infty} \frac{\delta_k \beta_{min}^{\rho+k} \gamma(\rho+k, \omega_d/\beta_{min})}{\Gamma(\rho+k)\beta_{min}^{\rho+k}} \\ &= \prod_{i=1}^{n} \left(\frac{\beta_{min}}{\beta_i}\right)^{\alpha_i} \sum_{k=0}^{\infty} \frac{\delta_k \gamma(\rho+k, \omega_d/\beta_{min})}{\Gamma(\rho+k)}. \end{aligned}$$

■

Proof of Proposition 4.5: The number of steps included in serial process i is defined by $\sum_{l=1}^{M} \theta_{i,l}$. Setting $n = \sum_{l=1}^{M} \theta_{i,l}$, the WTP of each pathway for time interval ω_d can be evaluated by $G(\omega_d)$ using Lemma 7.3. Then the WTP can be calculated from the weighted sum of the CDFs of all the pathways. ■

Appendix C. Proofs of Chapter 5

Proof of Proposition 5.1: For ED, the arrival process is defined by $\lambda_{\text{ext,ed}}$. The ED outputs to discharge, ICU and ward are characterized according to routing probabilities, $p_{\text{ed,dc}}$, $p_{\text{ed,icu}}$ and $p_{\text{ed,wd}}$, respectively. The latter two become the arrivals for ICU and ward, i.e.,

$$\lambda_{\text{ed,icu}} = \lambda_{\text{ext,ed}} p_{\text{ed,icu}},$$

$$\lambda_{\text{ed,wd}} = \lambda_{\text{ext,ed}} p_{\text{ed,wd}}.$$

For ICU, arrivals include those from ED, $\lambda_{\text{ed,icu}}$, and from ward, $\lambda_{\text{wd,icu}}$. Thus

$$\begin{aligned}\lambda_{\text{icu}} &= \lambda_{\text{ed,icu}} + \lambda_{\text{wd,icu}} \\ &= \lambda_{\text{ext,ed}} p_{\text{ed,icu}} + \lambda_{\text{wd,icu}}.\end{aligned}$$

As one can see, the latter is dependent on the departure from ward. For ward, arrivals include external admissions and transferring from both ED and ICU, i.e.,

$$\begin{aligned}\lambda_{\text{wd}} &= \lambda_{\text{ext,wd}} + \lambda_{\text{ed,wd}} + \lambda_{\text{icu,wd}} \\ &= \lambda_{\text{ext,wd}} + \lambda_{\text{ed}} p_{\text{ed,wd}} + \lambda_{\text{icu}} p_{\text{icu,wd}}.\end{aligned}$$

This leads to

$$\begin{aligned}
\lambda_{\text{wd,icu}} &= \lambda_{\text{wd}} p_{\text{wd,icu}} \\
&= (\lambda_{\text{ext,wd}} + \lambda_{\text{ed}} p_{\text{ed,wd}} + \lambda_{\text{icu}} p_{\text{icu,wd}}) p_{\text{wd,icu}} \\
&= \lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&\quad + \lambda_{\text{icu}} p_{\text{icu,wd}} p_{\text{wd,icu}} \\
&= \lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&\quad + (\lambda_{\text{ext,ed}} p_{\text{ed,icu}} + \lambda_{\text{wd,icu}}) p_{\text{icu,wd}} p_{\text{wd,icu}} \\
&= \lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&\quad + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}} p_{\text{wd,icu}} \\
&\quad + \lambda_{\text{wd,icu}} p_{\text{icu,wd}} p_{\text{wd,icu}},
\end{aligned}$$

which implies that

$$\begin{aligned}
\lambda_{\text{wd,icu}} = [&\lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&+ \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}} p_{\text{wd,icu}}] \\
&\cdot [1 - p_{\text{icu,wd}} p_{\text{wd,icu}}]^{-1}.
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
\lambda_{\text{icu}} &= [\lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&\quad + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}} p_{\text{wd,icu}}] \\
&\quad \cdot [1 - p_{\text{icu,wd}} p_{\text{wd,icu}}]^{-1} + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} \\
&= [\lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\
&\quad + \lambda_{\text{ext,ed}} p_{\text{ed,icu}}] \cdot [1 - p_{\text{icu,wd}} p_{\text{wd,icu}}]^{-1}.
\end{aligned}$$

It follows that

$$\begin{aligned}\lambda_{\text{wd}} &= \lambda_{\text{ext,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} \\ &\quad + [\lambda_{\text{ext,wd}} p_{\text{wd,icu}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} \\ &\quad + \lambda_{\text{ext,ed}} p_{\text{ed,icu}}] \cdot [1 - p_{\text{icu,wd}} p_{\text{wd,icu}}]^{-1} \cdot p_{\text{icu,wd}} \\ &= \frac{\lambda_{\text{ext,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}}}{1 - p_{\text{icu,wd}} p_{\text{wd,icu}}}.\end{aligned}$$

In addition,

$$\begin{aligned}\lambda_{\text{icu,wd}} &= \lambda_{\text{icu}} p_{\text{icu,wd}} \\ &= [\lambda_{\text{ext,wd}} p_{\text{wd,icu}} p_{\text{icu,wd}} + \lambda_{\text{ext,ed}} p_{\text{ed,icu}} p_{\text{icu,wd}} \\ &\quad + \lambda_{\text{ext,ed}} p_{\text{ed,wd}} p_{\text{wd,icu}} p_{\text{icu,wd}}] \\ &\quad \cdot [1 - p_{\text{icu,wd}} p_{\text{wd,icu}}]^{-1}.\end{aligned}$$

■

Appendix D. Proofs of Chapter 6

The proof of Proposition 6.2 requires Lemma 7.4.

Lemma 7.4 *Given Assumption 6.5, for any $f(s)$ that is nondecreasing in s, the following holds:*

$$\textstyle\sum_{s' \in \mathbb{S}} p_t(s'|s) f(s') \leq \sum_{s' \in \mathbb{S}} p_{t+1}(s'|s) f(s').$$

Proof of Lemma 7.4: By the definition of probability function, the following holds true:

$$\sum_{s'=0}^{S} p_t(s'|s) = \sum_{s'=0}^{S} p_{t+1}(s'|s) = 1.$$

Let $A = \sum_{s'=0}^{S} p_{t+1}(s'|s) - \sum_{s'=0}^{S} p_t(s'|s) = 0$. Then,

$$\begin{aligned} A &= \Big\{ \sum_{s'=0}^{S} p_{t+1}(s'|s) - \sum_{s'=0}^{S} p_t(s'|s) \Big\} f(0) \\ &\leq p_{t+1}(0|s)f(0) - p_t(0|s)f(0) + \Big\{ \sum_{s'=1}^{S} p_{t+1}(s'|s) - \sum_{s'=1}^{S} p_t(s'|s) \Big\} f(1) \\ &\leq \sum_{s'=0}^{1} p_{t+1}(s'|s)f(s') - \sum_{s'=0}^{1} p_t(s'|s)f(s') + \Big\{ \sum_{s'=2}^{S} p_{t+1}(s'|s) - \sum_{s'=2}^{S} p_t(s'|s) \Big\} f(2), \end{aligned}$$

since $f(s)$ is a nondecreasing function in s.

Continue in a similar fashion until we arrive at the following:

$$A \leq \sum_{s'=0}^{S} p_{t+1}(s'|s)f(s') - \sum_{s'=0}^{S} p_t(s'|s)f(s').$$

Since by definition $A = 0$, we have

$$0 \leq \sum_{s'=0}^{S} p_{t+1}(s'|s)f(s') - \sum_{s'=0}^{S} p_t(s'|s)f(s').$$

Hence,

$$\sum_{s'=0}^{S} p_t(s'|s)f(s') \leq \sum_{s'=0}^{S} p_{t+1}(s'|s)f(s').$$

■

Proof of Proposition 6.2: We use backward induction to prove the proposition.

Step 1: Show true for $t = T$.

$$v_T(s) = r_T(s, a) \geq 0 \geq v_{T-1}(s).$$

For $t \in \{0, 1, ..., T-1\}$, $v_t(s)$ only consists of costs, hence $v_t(s) \leq 0$ always holds true.

Step 2: Suppose true for $t = k + 1$.

$$v_{k+1}(s) \geq v_k(s).$$

Step 3: Show true for $t = k$.

$$
\begin{aligned}
v_k(s) &= max\Big\{r_k(s,a) + \textstyle\sum_{s' \in S} p_k^a(s'|s) v_{k+1}(s')\Big\} \\
&\geq max\Big\{r_{k-1}(s,a) + \sum_{s' \in S} p_{k-1}^a(s'|s) v_{k+1}(s')\Big\} \qquad \text{(D.1)} \\
&\geq max\Big\{r_{k-1}(s,a) + \sum_{s' \in S} p_{k-1}^a(s'|s) v_k(s')\Big\} \qquad \text{(D.2)} \\
&= v_{k-1}(s),
\end{aligned}
$$

where (D.1) holds due to Assumption 6.4, Proposition 6.1 and Lemma 7.4, while (D.2) holds true due to induction hypothesis that $v_{k+1}(s) \geq v_k(s)$. Hence, $v_{t+1}(s) \geq v_t(s)$ holds true for all t. ■

Bibliography

[1] David Scheinker and Margaret L Brandeau. Analytical approaches to operating room management. In *International Conference on Health Care Systems Engineering*, pages 17–26. Springer, 2017.

[2] James R Langabeer and Jeffrey Helton. *Health Care Operations Management: A Systems Perspective*. Jones & Bartlett Learning Burlington, MA, 2016.

[3] Claudia Steinke. Examining the role of service climate in health care: An empirical study of emergency departments. *International Journal of Service Industry Management*, 19(2):188–209, 2008.

[4] Victor J Dzau, Mark B McClellan, J Michael McGinnis, Sheila P Burke, Molly J Coye, Angela Diaz, Thomas A Daschle, William H Frist, Martha Gaines, Margaret A Hamburg, et al. Vital directions for health and health care: priorities from a national academy of medicine initiative. *Jama*, 317(14):1461–1470, 2017.

[5] Andrew B Rosenkrantz, Gregory N Nicola, Bibb Allen Jr, Danny R Hughes, and Joshua A Hirsch. Macra, mips, and the new medicare quality payment program: an update for radiologists. *Journal of the American College of Radiology*, 14(3):316–323, 2017.

[6] Randolph Hall. Patient flow. *AMC*, 10:12, 2013.

[7] Access is the answer: community health centers, primary care and the future of american health care. *National Association of Community Health Centers*, 2014.

[8] Scott A Shipman and Christine A Sinsky. Expanding primary care capacity by reducing waste and improving the efficiency of care. *Health Affairs*, 32(11):1990–1997, 2013.

[9] Thomas Bodenheimer and Hoangmai H Pham. Primary care: current problems and proposed solutions. *Health Affairs*, 29(5):799–805, 2010.

[10] X Zhong, M Williams, J Li, S Kraft, and J Sleeth. Primary care redesign: Review and a simulation study at a pediatric clinic. *Healthcare Analytics: From Data to Knowledge to Healthcare Improvement*, pages 399–426, 2016.

[11] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA: a cancer journal for clinicians*, 67(1):7–30, 2017.

[12] Ahmedin Jemal, Melissa M Center, Carol DeSantis, and Elizabeth M Ward. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiology and Prevention Biomarkers*, pages 1055–9965, 2010.

[13] Michael Williams. Hospitals and clinical facilities, processes and design for patient flow. In *Patient Flow: Reducing Delay in Healthcare Delivery*, pages 45–77. Springer, 2006.

[14] Barbara Resta, Vittorio Giudici, Sergio Cavalieri, Wei Deng Solvang, Stefano Dotti, and Paolo Gaiardelli. Strategic operations management in healthcare: A reference model for cardiac rehabilitation. In *International Conference on Health Care Systems Engineering*, pages 37–47. Springer, 2017.

[15] Stephen Trzeciak and Emanuel P Rivers. Emergency department overcrowding in the united states: an emerging threat to patient safety and public health. *Emergency medicine journal*, 20(5):402–405, 2003.

[16] Lucienne TQ Cardoso, Cintia MC Grion, Tiemi Matsuo, Elza HT Anami, Ivanil AM Kauss, Ludmila Seko, and Ana M Bonametti. Impact of delayed admission to intensive care units on mortality of critically ill patients: a cohort study. *Critical care*, 15(1):R28, 2011.

[17] Daniel A Handel, Joshua A Hilton, Michael J Ward, Elaine Rabin, Frank L Zwemer, Jr, and Jesse M Pines. Emergency department throughput, crowding, and

financial outcomes for hospitals. *Academic Emergency Medicine*, 17(8):840–847, 2010.

[18] Gregory Dobson, Hsiao-Hui Lee, and Edieal Pinker. A model of icu bumping. *Operations research*, 58(6):1564–1576, 2010.

[19] David W Manning, Adam I Edelstein, and Hasham M Alvi. Risk prediction tools for hip and knee arthroplasty. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 24(1):19–27, 2016.

[20] LB Moon and Jane Backer. Relationships among self-efficacy, outcome expectancy, and postoperative behaviors in total joint replacement patients. *Orthopedic nursing*, 19(2):77–85, 2000.

[21] Steven Kurtz, Kevin Ong, Edmund Lau, Fionna Mowat, and Michael Halpern. Projections of primary and revision hip and knee arthroplasty in the united states from 2005 to 2030. *JBJS*, 89(4):780–785, 2007.

[22] Julia Wong and Shirley Wong. A randomized controlled trial of a new approach to preoperative teaching and patient compliance. *International Journal of Nursing Studies*, 22(2):105–115, 1985.

[23] G Abelseth, RE Buckley, GE Pineo, Russell Hull, and M Sarah Rose. Incidence of deep-vein thrombosis in patients with fractures of the lower extremity distal to the hip. *Journal of orthopaedic trauma*, 10(4):230–235, 1996.

[24] Ville Remes. Corrinsights®: Are there modifiable risk factors for hospital readmission after total hip arthroplasty in a us healthcare system? *Clinical Orthopaedics and Related Research®*, 473(11):3456–3457, 2015.

[25] Stephen Yu, Kevin L Garvin, William L Healy, Vincent D Pellegrini Jr, and Richard Iorio. Preventing hospital readmissions and limiting the complications associated with total joint arthroplasty. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons*, 23(11):e60–e71, 2015.

[26] Nicholas J Vaudreuil, Timothy J McGlaston, Catarina D Gulledge, Allyn M Bove, and Brian A Klatt. Performance milestones in postoperative physical therapy after total hip arthroplasty: impact on length of stay and discharge destination. *Current orthopaedic practice*, 29(4):308–315, 2018.

[27] Marie D Westby and Catherine L Backman. Patient and health professional views on rehabilitation practices and outcomes following total hip and knee arthroplasty for osteoarthritis: a focus group study. *BMC health services research*, 10(1):119, 2010.

[28] Carlos J Lavernia, Michele R D'Apuzzo, Victor H Hernandez, David J Lee, and Mark D Rossi. Postdischarge costs in arthroplasty surgery. *The Journal of arthroplasty*, 21(6):144–150, 2006.

[29] Karen A Mauer, Elizabeth B Abrahams, Chris Arslanian, Lorry Schoenly, and Helen M Taggart. National practice patterns for the care of the patient with total joint replacement. *Orthopaedic Nursing*, 21(3):37–47, 2002.

[30] NIH et al. National institutes of health consensus statement on total knee replacement. *Bethesda (MD): US Department of Health and Human Services*, pages 1–18, 2004.

[31] Carlos J Lavernia, Victor Hugo Hernandez, and Mark D Rossi. Payment analysis of total hip replacement. *Current Opinion in Orthopaedics*, 18(1):23–27, 2007.

[32] Christine I Nichols and Joshua G Vose. Clinical outcomes and costs within 90 days of primary or revision total joint arthroplasty. *The Journal of arthroplasty*, 31(7):1400–1406, 2016.

[33] Paul Bower, S Campbell, Chris Bojke, and Bonnie Sibbald. Team structure, team climate and the quality of care in primary care: an observational study. *BMJ Quality & Safety*, 12(4):273–279, 2003.

[34] Chris Salisbury. Does advanced access work for patients and practices?, 2004.

[35] Mark Pickin, Alicia O'Cathain, Fiona C Sampson, and Simon Dixon. Evaluation of advanced access in the national primary care collaborative. *Br J Gen Pract*, 54(502):334–340, 2004.

[36] Louise Lemieux-Charles and Wendy L McGuire. What do we know about health care team effectiveness? a review of the literature. *Medical Care Research and Review*, 63(3):263–300, 2006.

[37] Allan H Goroll, Robert A Berenson, Stephen C Schoenbaum, and Laurence B Gardner. Fundamental reform of payment for adult primary care: comprehensive payment for comprehensive care. *Journal of General Internal Medicine*, 22(3):410–415, 2007.

[38] Thomas C Rosenthal. The medical home: growing evidence to support a new approach to primary care. *The Journal of the American Board of Family Medicine*, 21(5):427–440, 2008.

[39] Bruce E Landon, James M Gill, Richard C Antonelli, and Eugene C Rich. Prospects for rebuilding primary care using the patient-centered medical home. *Health Affairs*, 29(5):827–834, 2010.

[40] John W Beasley, Tosha B Wetterneck, Jon Temte, Jamie A Lapin, Paul Smith, A Joy Rivera-Rodriguez, and Ben-Tzion Karsh. Information chaos in primary care: implications for physician performance and patient safety. *The Journal of the American Board of Family Medicine*, 24(6):745–751, 2011.

[41] Michael E Porter, Erika A Pabo, and Thomas H Lee. Redesigning primary care: a strategic vision to improve value by organizing around patients? needs. *Health Affairs*, 32(3):516–525, 2013.

[42] James R Swisher and Sheldon H Jacobson. Evaluating the design of a family practice healthcare clinic using discrete-event simulation. *Health Care Management Science*, 5(2):75–88, 2002.

[43] Jared Reynolds, Zhen Zeng, Jingshan Li, and Shu-Yin Chiang. Design and analysis of a health care clinic for homeless people using simulations. *International journal of health care quality assurance*, 23(6):607–620, 2010.

[44] JR Villamizar, FC Coelli, WCA Pereira, and RMVR Almeida. Discrete-event computer simulation methods in the optimisation of a physiotherapy clinic. *Physiotherapy*, 97(1):71–77, 2011.

[45] Jacqueline Griffin, Shuangjun Xia, Siyang Peng, and Pinar Keskinocak. Improving patient flow in an obstetric unit. *Health care management science*, 15(1):1–14, 2012.

[46] Lixiang Jiang and Ronald E Giachetti. A queueing network model to analyze the impact of parallelization of care on patient cycle time. *Health Care Management Science*, 11(3):248–261, 2008.

[47] Linda Green. Queueing analysis in healthcare. In *Patient flow: reducing delay in healthcare delivery*, pages 281–307. Springer, 2006.

[48] Samuel Fomundam and Jeffrey W Herrmann. A survey of queuing theory applications in healthcare. Technical report, 2007.

[49] X Zhong, J Song, J Li, SM Ertl, and L Fielder. Analysis and design of gastroenterology (gi) clinic in digestive health center: A systems approach. *Flex Serv Manuf J*, 2015.

[50] Junwen Wang, Shichuan Quan, Jingshan Li, and Amy M Hollis. Modeling and analysis of work flow and staffing level in a computed tomography division of university of wisconsin medical foundation. *Health care management science*, 15(2):108–120, 2012.

[51] Xiang Zhong, Jingshan Li, Susan M Ertl, Carol Hassemer, and Lauren Fiedler. A system-theoretic approach to modeling and analysis of mammography testing process. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 46(1):126–138, 2016.

[52] Xiang Zhong, Jingshan Li, Philip A Bain, and Albert J Musa. Electronic visits in primary care: Modeling, analysis, and scheduling policies. *IEEE Transactions on Automation Science and Engineering*, 14(3):1451–1466, 2017.

[53] Junwen Wang, Xiang Zhong, Jingshan Li, and Patricia Kunz Howard. Modeling and analysis of care delivery services within patient rooms: A system-theoretic approach. *IEEE Transactions on Automation Science and Engineering*, 11(2):379–393, 2014.

[54] Pamela Mitchell and Robyn Golden. *Core principles & values of effective team-based health care*. National Academy of Sciences, 2012.

[55] Anni R Jensen, Jan Mainz, and Jens Overgaard. Impact of delay on diagnosis and treatment of primary lung cancer. *Acta Oncologica*, 41(2):147–152, 2002.

[56] Hardeep Singh, Kamal Hirani, Himabindu Kadiyala, Olga Rudomiotov, Traber Davis, Myrna M Khan, and Terry L Wahls. Characteristics and predictors of missed opportunities in lung cancer diagnosis: an electronic health record–based study. *Journal of Clinical Oncology*, 28(20):3307, 2010.

[57] N O'Rourke and R Edwards. Lung cancer treatment waiting times and tumour growth. *Clinical Oncology*, 12(3):141–144, 2000.

[58] Eija-Riitta Salomaa, Susanna Sällinen, Heikki Hiekkanen, and Kari Liippo. Delays in the diagnosis and treatment of lung cancer. *Chest*, 128(4):2282–2288, 2005.

[59] Dorothy S Lo, Robert A Zeldin, Roland Skrastins, Ian M Fraser, Harold Newman, Alan Monavvari, Yee C Ung, Harry Joseph, Teresa Downton, Larissa Maxwell,

et al. Time to treat: a system redesign focusing on decreasing the time from suspicion of lung cancer to diagnosis. *Journal of Thoracic Oncology*, 2(11):1001–1006, 2007.

[60] Feng Ju, Hyo Kyung Lee, Raymond U Osarogiagbon, Xinhua Yu, Nick Faris, and Jingshan Li. Computer modeling of lung cancer diagnosis-to-treatment process. *Translational lung cancer research*, 4(4):404, 2015.

[61] Cecily L Betz. Transition of adolescents with special health care needs: review and analysis of the literature. *Issues in comprehensive pediatric nursing*, 27(3):179–241, 2004.

[62] Mary D Naylor, Linda H Aiken, Ellen T Kurtzman, Danielle M Olds, and Karen B Hirschman. The importance of transitional care in achieving health reform. *Health affairs*, 30(4):746–754, 2011.

[63] JoAnna K Leyenaar, Arti D Desai, Q Burkhart, Layla Parast, Carol P Roth, Julie McGalliard, Jordan Marmet, Tamara D Simon, Carolyn Allshouse, Maria T Britto, et al. Quality measures to assess care transitions for hospitalized children. *Pediatrics*, page e20160906, 2016.

[64] Robert M Cowan and Stephen Trzeciak. Clinical review: emergency department overcrowding and the potential impact on the critically ill. *Critical care*, 9(3):291, 2004.

[65] Donald B Chalfin, Stephen Trzeciak, Antonios Likourezos, Brigitte M Baumann, and R Phillip Dellinger. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical care medicine*, 35(6):1477–1483, 2007.

[66] Fred Rincon, Stephan A Mayer, Juan Rivolta, Joshua Stillman, Bernadette Boden-Albala, Mitchell SV Elkind, Randolph Marshall, and Ji Y Chong. Impact of

delayed transfer of critically ill stroke patients from the emergency department to the neuro-icu. *Neurocritical care*, 13(1):75–81, 2010.

[67] Leora I Horwitz, Thom Meredith, Jeremiah D Schuur, Nidhi R Shah, Raghavendra G Kulkarni, and Grace Y Jenq. Dropping the baton: a qualitative analysis of failures during the transition from emergency department to inpatient care. *Annals of emergency medicine*, 53(6):701–710, 2009.

[68] Henry T Stelfox, Brenda R Hemmelgarn, Sean M Bagshaw, Song Gao, Christopher J Doig, Cheri Nijssen-Jordan, and Braden Manns. Intensive care unit bed availability and outcomes for hospitalized patients with sudden clinical deterioration. *Archives of internal medicine*, 172(6):467–474, 2012.

[69] Peter T Vanberkel, Richard J Boucherie, Erwin W Hans, Johann L Hurink, and Nelly Litvak. A survey of health care models that encompass multiple departments. *International Journal of Health Management and Information*, 1(1):37–69, 2010.

[70] Sara A Dolcetti. *Analyzing the impact of delays for patient transfers from the ICU to general care units*. PhD thesis, Massachusetts Institute of Technology, 2015.

[71] Yazan Roumani. *Modeling patient flow in a network of intensive care units (ICUs)*. PhD thesis, University of Pittsburgh, 2013.

[72] Hideaki Takagi, Yuta Kanai, and Kazuo Misue. Queueing network model for obstetric patient flow in a hospital. *Health care management science*, 20(3):433–451, 2017.

[73] Jeffery K Cochran and Aseem Bharti. Stochastic bed balancing of an obstetrics hospital. *Health care management science*, 9(1):31–45, 2006.

[74] Pouya Bastani. *A queueing model of hospital congestion*. PhD thesis, Dept. of Mathematics-Simon Fraser University, 2009.

[75] Naoru Koizumi, Eri Kuno, and Tony E Smith. Modeling patient flows using a queuing network with blocking. *Health care management science*, 8(1):49–60, 2005.

[76] Kurt M Bretthauer, H Sebastian Heese, Hubert Pun, and Edwin Coe. Blocking in healthcare operations: a new heuristic and an application. *Production and Operations Management*, 20(3):375–391, 2011.

[77] Thierry J Chaussalet, Haifeng Xie, and Peter H Millard. A closed queueing network approach to the analysis of patient flow in health care systems. *Methods of Information in Medicine*, 45(5):492–497, 2006.

[78] Carolina Osorio and Michel Bierlaire. An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. *European Journal of Operational Research*, 196(3):996–1007, 2009.

[79] Hyo Kyung Lee, Jingshan Li, Albert J Musa, Philip A Bain, and Kenneth Nelson. Modeling and analysis of patient transitions in community hospitals: A systems approach. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017.

[80] Remedios López-Liria, David Padilla-Góngora, Daniel Catalan-Matamoros, Patricia Rocamora-Pérez, Sagrario Pérez-de la Cruz, and Manuel Fernández-Sánchez. Home-based versus hospital-based rehabilitation program after total knee replacement. *BioMed research international*, 2015, 2015.

[81] Meghan E Daigle, Alexander M Weinstein, Jeffrey N Katz, and Elena Losina. The cost-effectiveness of total joint arthroplasty: a systematic review of published literature. *Best practice & research Clinical rheumatology*, 26(5):649–658, 2012.

[82] Ewa M Roos. Effectiveness and practice variation of rehabilitation after joint replacement. *Current opinion in rheumatology*, 15(2):160–162, 2003.

[83] Caroline Mitchell, Jane Walker, Stephen Walters, Anne B Morgan, Teena Binns, and Nigel Mathers. Costs and effectiveness of pre-and post-operative home physiotherapy for total knee replacement: randomized controlled trial. *Journal of Evaluation in Clinical Practice*, 11(3):283–292, 2005.

[84] Kate L Tribe, Helen M Lapsley, Marita J Cross, Brett G Courtenay, Peter M Brooks, and Lyn M March. Selection of patients for inpatient rehabilitation or direct home discharge following total joint replacement surgery: a comparison of health status and out-of-pocket expenditure of patients undergoing hip and knee arthroplasty for osteoarthritis. *Chronic illness*, 1(4):289–302, 2005.

[85] Nizar N Mahomed, Aileen M Davis, Gillian Hawker, Elizabeth Badley, J Rod Davey, Khalid A Syed, Peter C Coyte, Rajiv Gandhi, and James G Wright. Inpatient compared with home-based rehabilitation following primary unilateral total hip or knee replacement: a randomized controlled trial. *JBJS*, 90(8):1673–1680, 2008.

[86] Gerben DeJong, Ching-Hui Hsieh, Julie Gassaway, Susan D Horn, Randall J Smout, Koen Putman, Roberta James, Michael Brown, Elizabeth M Newman, and Mary P Foley. Characterizing rehabilitation services for patients with knee and hip replacement in skilled nursing facilities and inpatient rehabilitation facilities. *Archives of physical medicine and rehabilitation*, 90(8):1269–1283, 2009.

[87] Catherine J Minns Lowe, Karen L Barker, Michael E Dewey, and Catherine M Sackley. Effectiveness of physiotherapy exercise following hip arthroplasty for osteoarthritis: a systematic review of clinical trials. *BMC musculoskeletal disorders*, 10(1):98, 2009.

[88] Janet K Freburger. An analysis of the relationship between the utilization of physical therapy services and outcomes of care for patients after total hip arthroplasty. *Physical therapy*, 80(5):448–458, 2000.

[89] Tosan Okoro, Andrew B Lemmey, Peter Maddison, and John G Andrew. An appraisal of rehabilitation regimes used for improving functional outcome after total hip replacement surgery. *Sports Medicine, Arthroscopy, Rehabilitation, Therapy & Technology*, 4(1):5, 2012.

[90] Oguzhan Alagoz, Lisa M Maillart, Andrew J Schaefer, and Mark S Roberts. The optimal timing of living-donor liver transplantation. *Management Science*, 50(10):1420–1430, 2004.

[91] Jagpreet Chhatwal, Oguzhan Alagoz, and Elizabeth S Burnside. Optimal breast biopsy decision-making based on mammographic features and demographic factors. *Operations research*, 58(6):1577–1591, 2010.

[92] Jingyu Zhang, Brian T Denton, Hari Balasubramanian, Nilay D Shah, and Brant A Inman. Optimization of prostate biopsy referral decisions. *Manufacturing & Service Operations Management*, 14(4):529–547, 2012.

[93] Jingshan Li and Semyon M Meerkov. *Production systems engineering*. Springer Science & Business Media, 2008.

[94] Xiaolei Xie, Jingshan Li, Colleen H Swartz, and Paul DePriest. Improving the performance in acute care delivery: A systems approach. *IEEE Transactions on Automation Science and Engineering*, 11(4):1240–1249, 2014.

[95] Junwen Wang, Jingshan Li, and Patricia K Howard. A system model of work flow in the patient room of hospital emergency department. *Health care management science*, 16(4):341–351, 2013.

[96] Xiufeng Shao, Xiang Zhong, Jingshan Li, Bruce L Gewertz, Ken Catchpole, Eric J Ley, Jennifer Blaha, and Douglas A Wiegmann. Bottleneck analysis to reduce surgical flow disruptions: Theory and application. *IEEE Transactions on Automation Science and Engineering*, 12(1):127–139, 2015.

[97] Xiaolei Xie, Jingshan Li, Colleen H Swartz, and Paul DePriest. Modeling and analysis of rapid response process to improve patient safety in acute care. *IEEE Transactions on Automation Science and Engineering*, 9(2):215–225, 2012.

[98] Hyo Kyung Lee, Xiang Zhong, Jingshan Li, Albert J Musa, and Philip A Bain. Joint visit in primary care clinics: Modeling, analysis, and an application study. *IISE Transactions on Healthcare Systems Engineering*, 8(2):93–109, 2018.

[99] Offer Kella and Uri Yechiali. Waiting times in the non-preemptive priority m/m/c queue. *Stochastic Models*, 1(2):257–262, 1985.

[100] Randolph W Hall. *Queueing methods: for services and manufacturing*. Pearson College Div, 1991.

[101] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[102] Benjamin Zmistowski, Camilo Restrepo, Jordan Hess, Darius Adibi, Soltan Cangoz, and Javad Parvizi. Unplanned readmission after total joint arthroplasty: rates, reasons, and risk factors. *JBJS*, 95(20):1869–1876, 2013.

[103] Valerie Seagroatt, Heng Soon Tan, Michael Goldacre, Christopher Bulstrode, Ian Nugent, and Leicester Gill. Elective total hip replacement: incidence, emergency readmission rate, and postoperative mortality. *Bmj*, 303(6815):1431–1435, 1991.

[104] Samantha Tayne, Christian A Merrill, Eric L Smith, and William C Mackey. Predictive risk factors for 30-day readmissions following primary total joint arthroplasty and modification of patient management. *The Journal of arthroplasty*, 29(10):1938–1942, 2014.

[105] Nicholas L Ramos, Raj J Karia, Lorraine H Hutzler, Aaron M Brandt, James D Slover, and Joseph A Bosco. The effect of discharge disposition on 30-day readmission rates after total joint arthroplasty. *The Journal of arthroplasty*, 29(4):674–677, 2014.

[106] Stefano A Bini, Donald C Fithian, Liz W Paxton, Monti X Khatod, Maria C Inacio, and Robert S Namba. Does discharge disposition after primary total joint arthroplasty affect readmission rates? *The Journal of arthroplasty*, 25(1):114–117, 2010.

[107] Richard E Barlow and Frank Proschan. *Mathematical theory of reliability*, volume 17. Siam, 1996.

[108] Yuncheol Kang, Amy M Sawyer, Paul M Griffin, and Vittaldas V Prabhu. Modelling adherence behaviour for the treatment of obstructive sleep apnoea. *European journal of operational research*, 249(3):1005–1013, 2016.

[109] William W Schairer, David C Sing, Thomas P Vail, and Kevin J Bozic. Causes and frequency of unplanned hospital readmission after total hip arthroplasty. *Clinical Orthopaedics and Related Research®*, 472(2):464–470, 2014.

[110] Maria A Lopez-Olivo, Glenn C Landon, Sherwin J Siff, David Edelstein, Chong Pak, Michael A Kallen, Melinda Stanley, Hong Zhang, Kausha C Robinson, and Maria E Suarez-Almazor. Psychosocial determinants of outcomes in knee replacement. *Annals of the rheumatic diseases*, 70(10):1775–1781, 2011.

[111] Kirsten Jack, Sionnadh Mairi McLean, Jennifer Klaber Moffett, and Eric Gardiner. Barriers to treatment adherence in physiotherapy outpatient clinics: a systematic review. *Manual therapy*, 15(3):220–228, 2010.

[112] Jomon Aliyas Paul, Santhosh K George, Pengfei Yi, and Li Lin. Transient modeling in simulation of hospital operations for emergency response. *Prehospital and disaster medicine*, 21(4):223–236, 2006.

[113] Hyo Kyung Lee, Rebecca C Jin, Feng Yuan, Philip A Bain, Jo Goffinet, Christine Baker, and Jingshan Li. An analytical framework for tjr readmission prediction and cost-effective intervention. *IEEE journal of biomedical and health informatics*, 2018.

[114] Peter G Moschopoulos. The distribution of the sum of independent gamma random variables. *Annals of the Institute of Statistical Mathematics*, 37(1):541–544, 1985.