PRIMARY CARE DELIVERY SYSTEM: MODELING, ANALYSIS AND REDESIGN

By

Xiang Zhong

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

2016

Date of final oral examination: 06/01/2016

The dissertation is approved by the following members of the Final Oral Committee:
Jingshan Li, Professor, Department of Industrial and Systems Engineering
Ananth Krishnamurthy, Associate Professor, Department of Industrial and Systems Engineering
Douglas Wiegmann, Associate Professor, Department of Industrial and Systems Engineering
Marlon Mundt, Assistant Professor, Department of Family Medicine and Community Health
Zhengjun Zhang, Professor, Department of Statistics

Acknowledgements

My first sincere appreciation goes to my advisor Prof. Jingshan Li, who led me into the world of research and provided me tremendous support throughout my graduate study at the University of Wisconsin - Madison. His incredible mentorship has made a great deal of difference in my life. I am extremely fortunate to have him as my advisor.

I also heartfully thank my committee members: Prof. Marlon Mundt, Prof. Douglas Wiegmann, Prof. Ananth Krishnamurthy, and Prof. Zhengjun Zhang, not only for their insightful comments and encouragement, but also for the hard questions and challenges that incented me to widen my research from various perspectives.

I would like to thank Dr. K.P. Unnikrishnan and Dr. Jingyu Zhang, who provided me opportunities to join their teams as an intern at the NorthShore University HealthSystem and Philips Research North America, respectively.

During my Ph.D. study, I am lucky to have my colleagues Cong Zhao, Dr. Feng Ju, Dr. Xiaolei Xie, and Hyo Kyung Lee, who are also great friends to me. We have had a joyful time together, which made my days in Madison a warm and happy memory.

Lastly, I am grateful to my parents for their unconditional love and deep care, which always give me the strength to face any challenges in my life. This dissertation is lovingly dedicated to them.

Abstract

Primary care, the backbone of the nation's health care system, is at a grave risk of collapse and facing a confluence of factors that could spell disaster. Patients are dissatisfied and have difficulty of getting timely access while physicians are unhappy with their jobs by facing insurmountable tasks. There exists a strong call for improving patients' accessibility to care and enhancing providers' operational efficiency.

Analyzing patient flow plays an important role in improving the performance of health care delivery systems. Patient's length of visit (LOV), which characterizes the duration of an episode of hospital or clinic stay, is an extensively used measure to quantify the system performance. For traditional primary care visits, the care delivery services featuring multiple tasks accomplished by a limited number of care providers, comprise the majority of patient's LOV. It is critical for scientifically sound and valid methods to be developed and employed to capture the complexity of health care delivery systems and evaluate patient's LOV. However, few analytical work exists to study such processes with necessary details when surveying the literature. Thus, developing effective analytical models to study care delivery processes inside clinics is an objective of this research.

Meanwhile, the rapid development of information technology has introduced substantial opportunities and challenges in redesigning primary care. The advances in internet and mobile devices have made delivery of care over a distance possible. Many health care organizations have introduced online programs, referred to as "e-visits" (or "e-service", "e-portal", etc.), to provide patient-physician communication through securing messages. A spate of qualitative studies have investigated electronic messaging as a way to improve efficiency by decreasing the number of office visits. To better understand and implement e-visits, a mathematical model of primary care clinics with e-visits can provide a fresh look at the care delivery process from an integrated systems engineering perspective. Yet, few quantitative model is available in the current literature addressing e-visits in primary care. Therefore, another goal in this study is to establish an analytical framework for modeling primary care delivery with e-visits and investigate the impact of e-visits on care provider's productivity and patient's accessibility.

To achieve these goals, we start with modeling care service operations within patient rooms. A stochastic modeling framework is introduced to describe the workflow in outpatient clinics such as primary care and pediatric clinics, Gastroenterology (GI) clinics, and is also applicable to model hospital emergency departments and urgent care units. Furthermore, to resolve the dimensionality issue when extending the modeling scope to large-scale systems with shared resources, an iterative method, referred to as the shared resource iteration is proposed. Services within one exam room are modeled using the aforementioned approach and a convergent iterative method is applied to analyze the systems with two or more exam rooms.

On the other hand, to investigate e-visits' impact on primary care delivery, a queueing framework to study primary care physicians' operations coordinating patients' office visits, e-visits, and other non-direct care tasks is established. Analytical formulas to evaluate the average patient lengths of visit and their variances for both office visits and e-visits are derived. System monotonic properties are investigated and the conditions of when e-visits can lead to an improved access are identified.

Finally, to illustrate the applicability of the modeling scheme, case studies at the

GI clinic of the University of Wisconsin Health Digestive Health Center (DHC) and the Breast Imaging Center of the University of Wisconsin Medical Foundation (UWMF) are presented. Ways to improve the operational efficiency and to accommodate the rising patient demand are identified. The rigorous models and methods introduced in this dissertation provide quantitative tools for care providers to apprehend care delivery operations and design effective care delivery policies.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Motivation

The primary care delivery system is under tremendous strain. Due to population growth and aging, and the expanded health care insurance coverage, demand for primary care services has increased substantially in the past years. More patients need access to primary care but less medical students are choosing to enter the field. The recent studies show that 62 million people nationwide have no or inadequate access to primary care [1, 2] while only 13% of the final-year medical students are planning on primary care careers [3]. The implementation of the Affordable Care Act (ACA) will likely exacerbate the overcrowding in primary care clinics and the shortage of physicians [4]. Therefore, improving patients' accessibility to care and enhancing physicians' operational efficiency is of significant importance.

Analyzing patient flow in hospitals and clinics has been one of the key activities in the operations management of care delivery to improve system efficiency and patient satisfaction. A substantial amount of research has been devoted to studying patient flow to reduce delays in hospitals and clinics and facilitate the redesign of care delivery [5, 6]. Care delivery systems possess the complex, variable, dynamic and multidimensional nature. A spate of mathematical methods emerged; however, most of those methods only adopt deterministic models, which are inadequate because of the necessity for models to reflect the system dynamics. Meanwhile, an accurate and reliable stochastic model of patient flow would enable hospital and clinic managers to understand the system behaviors. Consequently, it can be useful in improving system functionality, such as predicting the future activity's impact on the wards and accommodating forthcoming demand variations.

To characterize the patient flow in clinics and hospitals, discrete-event simulation (DES) has acted as a prevailing approach for quantitative studies (e.g., reviews [7, 8, 9] and papers [10, 11, 12, 13, 14, 15, 16, 17, 18, 19]). Although simulations can provide a detailed and sometimes graphical representation, most of the simulation studies are case based. In other words, even if specific recommendations may be provided, system properties may not be unveiled. In addition, simulations typically require substantial details of data inputs and take a long time for model development. Moreover, in order to obtain statistically significant results, multiple replications are necessary and extended simulation time is ineluctable. These facts have limited the applicability of simulations. Analytical models, although less detailed compared to simulations, if with an appropriate level of simplification and abstraction, can provide a deft analysis of different system variations, and have the potential to uncover the insights and nature of the system. Therefore, analytical models are pursued in this dissertation to study the patient flow in health care delivery systems in an efficient and effective way. The following three subsections address the research motivations for primary care delivery system modeling, including modeling care activities within patient rooms, care delivery with shared resource, and systems incorporating e-visits.

1.1.1 Care Delivery within Patient Rooms

Care services within patient rooms are the most critical and time-consuming processes in primary care delivery, as well as other care delivery practices such as emergency departments. Most of the existing analytical methods to analyze patient flow use queueing theory models (e.g., review [20] and papers [21, 22, 23, 24]). Typically, such models use single or multiple servers to represent hospital or clinic operations at a very high level, without considering the details and complexities involved in the operations. For example, the specific activities within patient rooms during care delivery are typically ignored or aggregated into one operation using one server. Clearly, all the dynamics and behaviors within a patient room and their correlations with doctors and nurses are missing when they are represented by a single server. Therefore, lack of sophistication and fidelity constrains the application of using simple queueing theory models to characterize primary care delivery systems. To the best of our knowledge, few analytical models are available to describe the primary care services in detail. To fill the void in this area, Chapter 3 focuses on modeling care delivery services such as rooming, diagnosis, injection and immunization, and education within patient rooms. A Markov chain model featuring a closed, parallel, and reentrant network with limited resources is built to characterize the in-room care delivery process. Formulas to evaluate the patient length of visit and staff utilizations are developed, and the extension to non-Markovian scenarios is also investigated.

1.1.2 Care Delivery with Shared Resources

Although the method for modeling care delivery within patient rooms provides an approach to studying detailed care activities, substantial difficulties may arise when the resources (e.g. care providers, supportive staff) are shared among multiple activities. For instance, in primary clinics, laboratory test and diagnostic imaging centers, and some hospital units, the systems under study might be of large-scale, including multiple exam rooms and having complex staffing configurations. In addition, hospital units or clinics might not be able to invest sufficient capacity because of cost pressures, regulatory constraints, or a shortage of appropriate personnel. As a resolution, cross-trained staff with potentially increased service flexibility are adopted. However, although one supportive staff taking care of multiple exam rooms and patients could be capacity-wise saving, it also introduces availability issues among resources which could incur excessive waiting among patients.

Various care delivery systems share such identities and are desperate for effective methods to coordinate and improve care providers' operations. However, enormous challenges are encountered when shooting for a high fidelity model to capture complex system interactions. When characterizing a typical outpatient clinic using stochastic models, by just adding one room to the system, the size of the state space increases substantially, and the transitions are not easy to be unambiguously identified as they are subjected to the resource constraints. All these factors evince the plight in modeling complicated health care systems efficiently and effectively. Therefore, to resolve the dimensionality issue when modeling care delivery systems with resource sharing, an iterative method, referred to as the shared resource iteration, is introduced to characterize systems with one or several care providers taking care of multiple exam rooms and patients. In Chapter 4, a system-theoretic approach based on Markov chain models to study the patient flow of systems with resource sharing is introduced. The experimental results manifest that such a method results in an accurate performance estimation.

1.1.3 Care Delivery through Electronic Visits

The ever-evolving information technology has aroused substantial transformations in care delivery. The advances in internet and mobile devices have empowered patients to access care virtually. Electronic visits (or "e-visits", "e-portal", "e-service", etc.), where patient-physician communication is provided through securing messages, have attracted extensive attention from health professionals [4, 25, 26, 27, 28, 29, 30, 31, 32, 33]. Recent studies have demonstrated that significant savings can be obtained with improved access to care, increased provider efficiency and patient satisfaction, and lower costs by introducing e-visits, compared to traditional office visits [26, 33]. Electronic communication between patients and physicians can reduce health plan spending on physicians' office and laboratory services [34], and patients and physicians alike indicate satisfaction with electronic messaging [35, 36, 37]. However, the implementation of billable e-visits progresses slowly. Physicians hesitate to adopt e-visits for fears of being overburdened by electronic communication, and also the improper use of electronic communication by patients [28]. Many pilot studies on e-visits have been conducted through observations, interviews, and survey analyses. To better understand and implement e-visits, a mathematical model characterizing primary care delivery with e-visits can provide a fresh perspective of the care delivery process from an integrated systems engineering's point of view. As indicated in the 2014 Report to the President by Presidents Council of Advisors on Science and Technology [38], and the 2009 Agency for Healthcare Research and Quality (AHRQ) and National Science Foundation (NSF) workshop [39], such systematic approaches can lead to a breakthrough towards a new era of care delivery modeling.

To implement e-visits to increase care accessibility, a major issue is to understand how patient's access to care can be impacted by having more care options such as e-visits [25].

Specifically, questions such as how is the workflow in primary care clinics affected by the use of e-visits, and what is the impact on resources necessary to deliver proper care arise naturally. To answer these questions, evaluating the efficiency of primary care operations with e-visits, and determining the optimal scheduling policies coordinating office and e-visits are aspired. Unfortunately, the current literature lacks effective methods to address these problems. To bridge the gap, Chapters 5 and 6 are devoted to establishing an analytical framework for modeling the primary care operations with e-visits and using it to resolve the accessibility issue in primary care.

1.2 Organization of the Document

The rest of this document is organized as follows. Chapter 2 reviews the related literature on primary care redesign and the modeling techniques and methods applied in health care systems. Chapter 3 presents the analytical modeling of in-room care services. Formulas for performance evaluation are derived and system properties are examined. In Chapter 4, to address the dimensionality issue when modeling systems with multiple exam rooms and shared resources, an iterative method is introduced. The convergence of the method is justified analytically. This Markov chain based modeling framework generates accurate system performance estimates. To incorporate e-visits in primary care, Chapter 5 introduces a queueing framework to model physicians' operations coordinating e-visits and office visits, as well as physicians' non-direct care activities. The impact of e-visits on patient access to primary care is discussed in Chapter 6. Furthermore, to elucidate the applicability of the proposed modeling framework, a case study at the Gastroenterology clinic of the University of Wisconsin Health Digestive Health Center, and a case study of mammography testing process at the Breast Imaging Center of the University of Wisconsin Medical Foundation are introduced in Chapter 7. Finally, the summary and future work are presented in Chapter 8. All the proofs and derivations can be found in the Appendix.

Chapter 2

Literature Review

The goal of this research is to develop rigorous engineering approaches to model primary care delivery systems to improve patient accessibility and provider efficiency. Thus, the literature review focuses on the following aspects: Section 2.1 highlights the research efforts on redesigning primary care. Section 2.2 reviews the care activity modeling, the modeling methods and techniques and the applications in health care systems. The existing research on e-visits are summarized in Section 2.3. Finally, Section 2.4 illustrates the perspective of the current study.

2.1 Primary Care Redesign

In the face of the diminishing government subsidies, fierce competition, and the influence of care reform, health care organizations are rife with pressures to change [40, 41]. The primary care redesign initiative aims to provide easy and timely access to care, which is culturally sensitive, quality-driven, and maximizes the use of education and community resources based on patient needs. To make such a care delivery model successful, it is essential to create a sustainable environment, in which team members work to their highest level of licensure with excellence at all levels of the system of care, and make full use of existing and emerging technologies [42].

Redesigning primary care has attracted substantial research efforts (see white paper [43], and reviews [41, 44, 45, 46]). The national conundrums in primary care have been outlined by Bodenheimer [45], where a confluence of factors could lead to disasters, such as excessive demands, uneven quality of care, unhappiness with jobs, and inadequate reimbursement. To address these challenges, actions on primary care practices (microsystem improvement) and larger health care systems (macrosystem reform) are desired. Furthermore, Bodenheimer and Pham [46] review the state of primary care in the US and have conducted a thorough discussion on the feature and landscape of primary care practices. The difficulties accessing primary care are due to multiple factors, including shortage in the primary care practitioner workforce, geographic maldistribution, and organizational issues within primary care practices. Issues such as teamwork, electronic health records (EHRs) and information systems, medical homes, payment systems, as well as advanced access are of primary interests by researchers. To provide a general picture of primary care redesign, these studies are reviewed in Subsections 2.1.1 - 2.1.5, correspondingly.

2.1.1 Teamwork

Teamwork in primary care practice has proven benefits in achieving better outcomes. Lemieux-Charles and McGuire [47] provide a review of health care team effectiveness from 1985 to 2004 by comparing teams with usual (nonteam) care, examining the impact of team design on effectiveness, and exploring relationships between team context, structure, processes, and outcomes. Through observational studies, Bower et al. [48] discuss team practice structure, process (climate), and outcome (quality of care) in primary care. The results indicate that there exist important relationships between team structure, process, and outcome that may be of relevance to quality improvement initiatives in primary care. More studies on teamwork in primary care can be found in [49, 50, 51, 52, 53].

2.1.2 Electronic Health Records

Data in electronic health records plays a central role in care delivery, quality control, clinical governance, and provider practices. The adoption of EHR system has been a worldwide trend in care practice. Lusignan and Weel [54] review the opportunities of using routinely collected data in primary care research, such as handling growing volumes, improving data quality, facilitating technological progress for processing, and bridging clinical and genetic data, as well as establishing the body of know-how within health informatics community. A comprehensive review of the literature on the current state of the implementation of health information system in primary care is carried out by Ludwick and Doucette [55]. It highlights the factors that affect EHR implementation outcomes, which include the graphical user interface design quality, feature functionality, project management, procurement and users' previous experience. In addition, the data quality in electronic patient records (EPRs) in primary care is reviewed by Thiru et al. [56] based on publications in 1980 to 2001. Hillestad et al. [57] investigate the impact of electronic medical record (EMR) systems on transforming primary care, and the potential health benefits, savings, and costs. From a human-factors engineering perspective, Beasley et al. [58] discuss the concept of information chaos in primary care and explores the implications and impacts on physician performance and patient safety. More studies related to data and information systems in primary care are discussed in papers [59, 60, 61, 62].

2.1.3 Medical Home

The concept of medical home is defined as having a regular doctor or place of care, doctor/staff knowing information about patient's health history, the place being easy to contact by phone, and the doctor/staff coordinating care received from other doctors or source of care [63]. Rosenthal [64] reviews the literature and programs on medical homes to assess the usefulness of the model based on several principles, such as team-directed medical practice, personal physician, whole-person orientation, coordinated and integrated care across the health care domain, as well as quality and safety. By arguing that the specialist-dominated US health care system results in mediocre quality care with the excessive use of costly service but little marginal health benefit, Landon et al. [65] further claim that the patient-centered medical home has become a policy shorthand for rebuilding US primary care capacity. Papers [66, 67, 68, 69] provide more references addressing medical homes in primary care practice.

2.1.4 Payment System

Davis et al. [44] argue that a new primary care payment system to blend monthly patient panel fees, traditional fee-for-service, and new incentives for patient-centered care performance is desirable. As performance-based payments are increasingly common in primary care, Friedberg et al. [70] suggest that pay-for-performance programs should monitor and address the potential impact of performance-based payments on health care disparities. To improve the ability of primary care to play its essential role in the care delivery system, Porter et al. [71] offer a framework based on value for patients to sustain and improve primary care practice. It states that payment should be modified to bundle reimbursement for each subgroup and reward value improvement. Extensive studies in payment related issues in primary care have been introduced in papers [72, 73, 74, 75].

2.1.5 Advanced Access

The advanced access, also known as open access or same-day scheduling, in which patients calling to schedule a physician visit are offered an appointment on the same day, has manifested its helpfulness to reduce waiting times in primary care. Murray et al. [76, 77] summarize six elements of advanced access that make it sustainable: balancing supply and demand, reducing backlogs, reducing the variety of appointment types, developing contingency plans for unusual circumstances, working to adjust demand profiles, and increasing the availability of bottleneck resources. Surveys in papers [78, 79] also show that patients are seen more quickly in advanced access practices, but with less flexibility in the choice of appointment. Thus, appointment systems should be flexible to accommodate different needs of disparate patient groups. Additional papers studying advanced access in primary care can be found in [80, 81].

In addition to the five aspects discussed above, other reform strategies are also proposed to address the problems of estimating panel size, increasing capacity, and mitigating geographic maldistribution. Other recommendations include standardizing reimbursement levels to reduce insurance-linked refusal, increasing after-hour access, implementing open access for same-day scheduling, introducing e-mail and telephone visits, and forming primary care teams with nonprofessional team members.

2.2 Care Operations Modeling

In response to the primary care redesign initiative, many health care organizations have made drastic changes yet resulted in slightly increased patient access without any significant reductions in cost. Moreover, improved clinical outcomes and service quality are increasingly demanded by payers and patients. All these factors combined with the complex and dynamic system nature make care delivery systems the area for the development and use of OR/MS methods and frameworks to help identify capacity needs and system bottlenecks. The scope of this review is to bring together the recent developments that are related to patient flow and care delivery operations modeling. Specifically, computer simulations and analytic models are the prevailing OR/MS methods and will be elaborated in Subsections 2.2.1 and 2.2.2, respectively.

2.2.1 Discrete-Event Simulation in Care Operations Modeling

Operational inefficiencies have their roots in the improper benchmarking and unaccounted dynamics. It is important for health systems to react to redesign on an ad hoc basis. Many health systems use "small tests of change" to iteratively improve processes. However, when contemplating major changes in systems of care, such as appointment scheduling and staffing, the use of the plan-do-check-act (PDCA) model is not appropriate for being disruptive and time-consuming. Meanwhile, simulation offers an alternative method to "test" changes in practice and to evaluate the impact of those changes on patients and staff. In recent years, discrete-event simulation dominates the quantitative studies in care delivery research (see reviews [8, 82, 83]). The booming information technology and data analytics have substantially enhanced and extended the functions of simulation tools. Through modeling complex facilities, sophisticated logics, and dynamic schedules, simulation has become the very aid for decision making and operations improvement. In particular, simulation models are widely used for assessing system efficacy, carrying out what-if analysis to evaluate the system design, studying the impact of potential changes, and investigating the complex relationships among system variables. A comprehensive review of discrete-event simulation in health care is presented by Jacobson et al. [8]. In this review, simulation studies of single or multi-facility healthcare organizations including outpatient clinics, emergency departments, surgical centers, orthopedic departments, and pharmacies are reviewed. Similar reviews have been provided in [82, 83, 84]. For instance, Gunal and Pidd [82] classify the papers of discrete-event simulation for performance modeling in health care according to the areas of applications. They indicate that there is a lack of generality and explain the rationale why generic approaches are rare and specificity dominates. In addition, by reviewing the legacies of simulation modeling in health care, Eldabi et al. [83] propose future opportunities to use simulation as a problem-solving technique in health care settings. Eldabi et al. point out that a major challenge lies in persuading service providers that simulation can make a critical contribution, and from the perspective of the modeling community, the most pressing need is to join up different modeling methods. Moreover, Wiler et al. [84] focus on the emergency department (ED) and categorize the modeling approach of patient flow in emergency departments into five categories: formula-based, regression-based, time series analysis, queueing models, and discrete-event simulations.

As ED is one of the most critical departments in a hospital, and ED overcrowding has become a national crisis ([85, 86]), a substantial amount of simulation studies have been devoted to ED to reduce crowding (see reviews [84, 87]). Additional simulation studies in EDs have been reported in papers [13, 14, 15, 88, 89, 90]. In addition to simulations of EDs, other hospital departments are also studied extensively. Simulation studies on different hospital units (e.g., critical care, intensive care, surgical, pharmacy, medical, etc.) can be found in [91, 92, 93, 94, 95, 96]. In addition, simulations have been used to inspect and improve the scheduling and appointment systems in outpatient clinics, see papers [97, 98, 99, 100, 101]. More simulation studies on outpatient clinics are described in [102, 103, 104, 105, 106].

2.2.2 Analytical Methods in Care Operations Modeling

While simulations are extensively applied and can provide detailed analysis, the majority of them are case study based and may suffer from long model development and simulation time. The increasing variance has also brought a mounting awareness of the limitations of conventional simulation techniques. Meanwhile, analytical tools, such as queueing theory models, can provide a quick analysis and have the potential to dig deeper into system properties. Compared with simulation models, analytical models have been used much less frequently to study primary care operations. Reviews of such models can be found in [20, 84, 107]. Fomundam and Herrmann [20] summarize queueing theory applications in health care, such as waiting time and utilization analysis, system design, and appointment systems comparisons at different scales, from individual departments (or units) to healthcare facilities and regional health care systems. Green et al. [21] introduce an M/M/s queueing model to estimate the number of providers needed in an emergency department. To determine bed capacity of maternity facilities in a perinatal network, a queueing theory model is used by Pehlivan et al. [108] to evaluate the refused admission probability. Such a model is embedded into a multi-period mixed-integer optimization algorithm for estimating the necessary capacity. Approximating generating function analysis is presented by Au-Yeung et al. [22] to study patient response time. Through using a queueing model, where the nodes represent assessment and treatment stages, patient completion times are analyzed by Mayhew and Smith [23]. Similarly, a queueing network model is described by Jiang and Giachetti [24] to investigate the impact of parallelization of care on patient cycle times. Dobson and Lee [109] develop a stochastic model of an ICU with patient bumping under differing capacity and arrival patterns. Abraham et al. [110] compare several models for forecasting daily emergency inpatient admissions and occupancy.

In addition to queueing models, other analytical methods have been introduced to conduct flexible analysis and gain insights. For example, a three-level strategy model to design a hospital department is presented by Fanti et al. [111] with three basic elements: 1) modeling module, 2) optimization module, and 3) simulation and decision module. Augusto and Xie [112] introduce a new modeling methodology to address organization problems of health care systems using Petri-nets based metamodels. The issues of outpatient appointment scheduling are studied in [113, 114, 115] using mixed-integer programming models.

Markov models are often used to represent stochastic processes which are formalized by a set of states to which the system may belong, and probabilistic laws that govern the movement between the states. Such models assume the probabilistic behavior of patients moving around the system and therefore, gives a realistic representation of the actual system. In an early paper, Irvine et al. [116] describe the development of a continuous time stochastic model of patient flow. Essentially, it is a two-stage continuous-time Markov model that describes the movement of patients through geriatric hospitals. Further, McClean et al. [117] extend stochastic Markov modeling to three stages and attach different costs to each of the three stages. Such a model can facilitate planning of health and social services for the elderly while taking cost into account. Taylor et al. [118] use a continuous time Markov model and apply it to a four compartmental model where the four stages are acute, long-stay, community, and dead. Then Taylor et al. [119] extend these models to contain six stages to determine the interactions between hospital geriatric medical services and community care. In Wang et al. [120], a Markov chain model is developed to analyze the workflow and staffing level in a CT division of the UWMF. In the study of the rapid response process to improve patient safety in acute care, the response process is modeled as a complex network with split, merge, and parallel structures and an analytical method is developed to evaluate the decision time and its variability [121, 122].

2.3 Electronic Visits in Primary Care

Amid the redesigned primary care, e-visits, as a novel alternative to the traditional office visits, starts to attract growing attention in recent years. Widespread efforts to improve health care quality, safety, and efficiency focus on using information technologies including electronic health records, patient registries, computerized physician order entry, and embedded decision supports. Among them, e-visits, involving the usage of patient portals where patients can access their medical records and communicate with their primary care providers by secure messaging, has been viewed as a promising technology to improve the quality and efficiency of care [123]. This innovative channel of care delivery has received considerable attention in the US. Many health care organizations, including Henry Ford Health System, Mayo Clinic, Kaiser Permanente Health Plan, and the University of Pittsburgh Medical Center have initiated e-visit programs [4, 25, 26, 31, 32, 33]. According to a survey by Manhattan Research (Wall Street Journal 2012 [124]), the percentage of physicians who say they use secure messaging, e-mail, instant messaging or video conferring with their patients has increased from below 25%

in 2005 to about 40% in 2011. Primary care providers eager for patients to adopt e-visits as the technology holds promise to reduce the physician workload on office visits and telephone visits on top of improving patient health. Most of the existing e-visit studies focus on investigating the effectiveness and patient/provider experiences of implementing e-visits. It is found that the quality of care and patient outcomes using e-visits are equivalent to those achieved with office visits [31, 32].

Implementing e-visits can free up extra office appointments for the patients with urgent and complicated issues, reduce urgent care and emergency room visits and inpatient hospital admissions, improve care for senior population with chronic diseases, and substantially reduce the cost of care [25, 26, 31, 32, 33]. However, there are mixed conclusions drawn regarding the substitutability of e-visits with traditional forms of physician contact. Specifically, Katz et al. [125, 126] suggest that e-visits have no discernible effect on reducing physician workload, and e-mails generated through a triage-based system did not appear to substitute for phone communication or to reduce visit no-shows in a primary care setting. In addition to that, other studies investigate billing and reimbursement issues, information system structures, legal and regulatory issues, financial return, and system implementation and training [4]. The empirical evaluation of e-visits is challenging and, therefore, quantitative models to study e-visits and its impact on primary care delivery systems are pursued. As a quantitative analysis of e-visits, a patient health dynamics model is developed in [127] under alternative primary care delivery modes, including the usage of e-visits and non-physician providers. This study quantifies the overall impact of adopting e-visits on physician's choices and expected earnings and patients' expected health outcomes. In a follow-up study based upon these results, it is argued that e-visits provide a gateway for traditional forms of primary care delivery

[128]. In spite of these efforts, no analytical study on patient flow and operations management has been carried out for primary care delivery through e-visits. Moreover, all the existing studies do not consider the detailed workflow design for physicians to handle the increasing amount of e-visits, and often overlook the scenario that the providers may not be available for clinical service due to other obligations.

2.4 Perspectives

Health care providers are increasingly aware of the need to use their resources as efficiently as possible to improve patients' ability to receive the most appropriate care in a timely fashion. As this chapter has attempted to demonstrate, effective engineering approaches are critical to this objective and have achieved cheerful accomplishments. Yet, complexities which feature various types of patients, time-varying demands, and the often disparate perspectives of administrators, physicians, nurses, and patients are embedded in care delivery systems. These pervasive challenges affect the ability and performance of the existing methods to improve the quality of care delivery. Care providers are still longing for effective methods to gain managerial insights and make decisions. The urge for analytical tools to study patient flow and enhance operations management for redesigning primary care is eminence, and our work intends to contribute to this end. In close, modeling patient flow in health care systems can assist in the overall understanding of the system activities and be useful in improving the system performance and functionality. The success of modeling care operations would contribute to the interdisciplinary research in redesigning primary care.

Chapter 3

Modeling of Care Delivery Operations within Patient Rooms

3.1 Introduction

In this chapter, we introduce a Markov chain model featuring a closed, parallel, and reentrant network with limited resources to investigate the in-room care delivery process. The system under study is described in Section 3.2. Formulas to evaluate the patient length of visit and staff utilizations are developed in Section 3.3. The extension to non-Markovian scenarios is investigated and approximation formulas are presented in Section 3.4.

3.2 System Description and Problem Formulation

3.2.1 System Description

The patient flow in an outpatient clinic typically includes the following processes: registration/triage, waiting for room assignment, in-room care services (physician visit, medical assistant visit, nurse education, intravenous (IV) administration, medical assistant warp-up, etc.), and finally check-out. Among those activities, the care services delivered within patient rooms are extremely critical since the majority of the value-added time is spent here. In addition, most of the resources or capacity of the clinic, such as physicians (MDs), medical assistants (MAs), nurses (RNs), technicians, and exam rooms contribute to this part. Therefore, accurate quantitative analysis of the patient flow and care delivery services within patient rooms is compelling.

By considering the services a patient may receive within the patient room, the patient flow in a typical outpatient clinic are described in Figure 3.1, where the circles represent the care services and the solid lines characterize the patient flow in terms of receiving these services. Each line represents an individual patient room and the corresponding care operations in the room. As exhibited in Figure 3.1, a patient starts with rooming with medical assistant, then physician diagnosis, and then nurse administering IV if necessary, physician revisit if necessary, and finally, medical assistant warp-up.

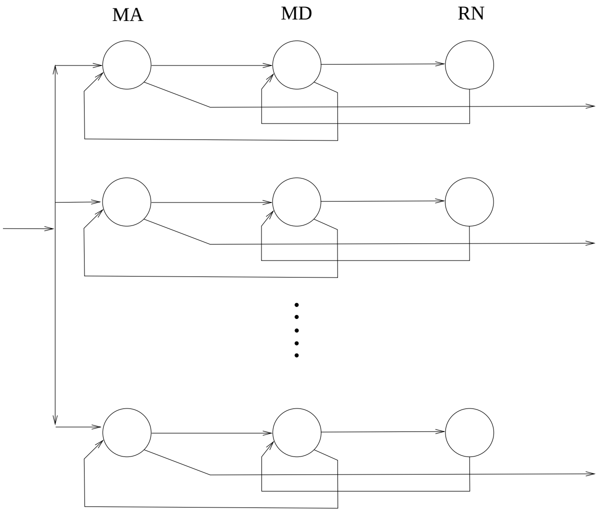


Figure 3.1: In-room patient flow in outpatient clinics

In addition to the general patient flow described above, it's possible that patients may not need IV administration and will exit directly. There are also cases that patients do not need to visit physicians or MAs for the second time. Therefore, the basic patient flow in the patient room can be represented by a closed re-entrant process with splits. The "re-entrant" characterizes repeated visits of physicians and MAs, the "splits" describes disparate patient flows (i.e., who may skip IVs and the subsequent services), and the patient room (or bed) represents the "closed" nature, i.e., the room can be viewed as a carrier to undergo all the services with the patient, and will be available to the next patient after the current one leaves. Typically, there could be multiple patients presented in multiple rooms simultaneously. Therefore, it is a complex system with multiple, parallel, and re-entrant processes with splits. Moreover, the resources (such as physicians, MAs, nurses, or equipment) are limited.

Remark 3.1 Note that the flow described in Figure 3.1 represents the care delivery services for patients during their stays. A patient may not physically present in the patient room throughout their visits. In some occasions, a patient might need to leave the room for other tests, but the room is typically reserved for the patient before check-out. In particular, in emergency departments, new patients will not be admitted into the room until the current one is discharged. Therefore, it is equivalent that the room "goes through" all the services with the patient.

3.2.2 Structural Modeling

Clearly, the system described in Figure 3.1 is too complex to solve directly. In order to reduce the complexity of the process and make it analytically tractable, we aggregate the split patient routes into one. In other words, assuming the patients who do not need IV administration and follow-up visits by physicians and MAs still follow the general flow of nurse-physician-MA, but with minimum (or zero) service times, we obtain a unified care delivery process in a patient room.

Nevertheless, the closed and re-entrant characteristics still remain and to further reduce the complexity, some special system features should be utilized. Notably, there is only one patient in each patient room (or bed) at one time. It implies that the population within a closed re-entrant process is normally one. In addition, upon finishing the service, the care provider will be released from this activity and the patient will wait for the next service (e.g., the MA leaves after carrying out the initial inquiry, and the patient needs to wait for the physician). This is equivalent to the patient waiting in a buffer for the next "operation." Therefore, such a closed re-entrant process can be represented by a closed serial line with population one, see Figure 3.2.

Remark 3.2 The serial processes characterize the sequential care services provided to patients during his/her stays. In some cases, patients may be served by physicians and MAs more than two times, and additional tests or other care services may be provided. Then, an equivalent serial process with multiple visits (more than twice) by physicians, MAs, or nurses can be constructed. Although consisting of more services and variations, similar approaches can be applied to processes of this type.

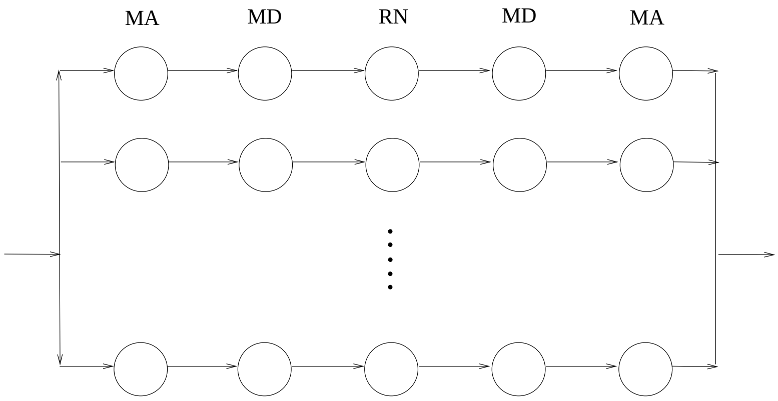


Figure 3.2: Equivalent parallel-serial processes of multiple patient rooms

3.2.3 Assumptions

To analyze the system as described in Subsection 3.2.1, the following assumptions are introduced to address the services, providers (resources), and their interactions.

(i) There are M patient rooms (or beds) available in an outpatient clinic, where each room (or bed) can accommodate one patient only.

(ii) The patient arrival is unconstrained. In other words, the patient room will be immediately occupied by a new patient right after the previous one is discharged. The unconstrained arrival assumption can be relaxed to general arrival distributions.

Remark 3.3 Ideally, patients should arrive at the clinic based on their appointment time. In reality, variations such as early or late arrival or even no-show cannot be overlooked. Poisson arrival, where the inter-arrival time of the incoming patients follows an exponential distribution are prevailingly used to describe the patient arrival process [129]. The method to handle more general arrival distributions will be discussed in Chapter 4. However, when the patient demand is high, there will always be patients ready to be roomed. The unconstrained arrival can be assumed.

(iii) There are N services, including physician's first and second visits, MA's initial visit and medication or wrap-up, nurse administering IV, etc. We assume that all patient rooms are identical, such that for each service (for example, MA's initial visit), the processing times are identical among all patient rooms, described by an exponential distribution with a mean processing time τ_i, $i = 1, \ldots, N$. Then, the processing rate of service i is $c_i = \frac{1}{\tau_i}$, $i = 1, \ldots, N$.

Remark 3.4 The services are carried out by the same group of providers. When patients have an equal probability to be assigned to any available room, the assumption of the identical processing time among all patient rooms is valid.

Remark 3.5 The exponential service time is introduced to make the analysis tractable. In Section 3.4, such an assumption will be relaxed and non-exponential service time distributions can be addressed.

(iv) There are R types of resources in the system, including physicians, MAs, nurses, etc. The quantity of each resource is defined by $r_j \in \{1, 2, \ldots, M\}$, $j = 1, \ldots, R$.

(v) Each care service can only be carried out when the required resource (e.g., the physician visit requires one physician) is available. Parameters $\theta_i \in \{1, 2, \ldots, R\}$, $i = 1, \ldots, N$, define the type of resource for service i, and $\theta_i = j$ indicates that resource type j is needed for service i.

Remark 3.6 To avoid messy notations, for the service that does not need any resource, we still assume the existence of a virtual resource, but with a quantity equal to the number of rooms M. This is equivalent to having available resources anytime.

(vi) In some cases, two or more services may require the same type of resource (e.g., both MA's initial visit and wrap-up need an MA as the resource). The priority of a service to grab the resource is defined by parameters $p_i \in [0, 1]$, $i = 1, \ldots, N$, where $p_i = 1$ represents the highest priority, and 0 the lowest. For simplicity, we assume none of these services have the same priority.

(vii) The services are delivered by their designated resources. When multiple services request the same resource at the same time, the resource will select the service based on the rank of priority.

3.2.4 Problem Formulation

In an appropriately defined state space, the system under assumptions (i)-(vii) formulates a stationary random process. Let T_s denote the patient length of visit in the patient room. In the framework of (i)-(vii), T_s is a function of all system parameters:

$$T_s = f_t(\mathcal{C}, \mathcal{P}, \mathcal{R}, \Theta, M), \tag{3.1}$$

where

$$\begin{aligned} \mathcal{C} &= [c_1, c_2, \ldots, c_N], \\ \mathcal{P} &= [p_1, p_2, \ldots, p_N], \\ \mathcal{R} &= [r_1, r_2, \ldots, r_R], \\ \Theta &= [\theta_1, \theta_2, \ldots, \theta_N]. \end{aligned}$$

Similarly, the utilization of resources (care providers and equipment), denoted as ρ_k, $k = 1, \ldots, R$, is also a function of all parameters,

$$\rho_k = f_\rho(\mathcal{C}, \mathcal{P}, \mathcal{R}, \Theta, M), \quad k = 1, \ldots, R. \tag{3.2}$$

The problem to be addressed is: *Under assumptions (i)-(vii), develop a method to evaluate the patient length of visit within patient rooms and staff utilization as functions of the system parameters.*

3.3 Performance Analysis

3.3.1 State Space

The states of the system can be represented by $S = (n_1, n_2, \ldots, n_N)$, where n_i represents the number of patients in service i, and $n_i \in \{0, 1, 2, \ldots, M\}$, $i = 1, \ldots, N$. Since at most one patient is allowed in each patient room, and due to the unconstrained arrival assumption, there're always M patients in the system:

$$\sum_{i=1}^{N} n_i = M. \tag{3.3}$$

With constraint (3.3), the effective or available states are reduced dramatically. To find the effective state space, let $f_M(m, j)$ denote the number of states in a M-room system if there are m patients in the first service and there are j services in each room. If all M patients are undertaking the first service, then all other services will have no patient, resulting in one state $S = (M, 0, \ldots, 0)$. Moreover, if there are m patients ($m < M$) in the first service, then the remaining $M - m$ patients are distributed in $j - 1$ services. This implies that there will be $M - m - k$ patients in $j - 2$ services if k patients are in the second service. By repeating this argument, we have $f_M(m, j)$ being calculated as follows:

$$\begin{aligned} f_M(m,1) &= 0, \quad m = 0, 1, \cdots, M-1, \\ f_M(M,j) &= 1, \quad j = 1, 2, \cdots, N, \\ f_M(m,j) &= \sum_{k=0}^{M-m} f_{M-1}(M-m-k, j-1), \\ &\qquad m = 0, 1, \cdots, M; j = 2, \cdots, N. \end{aligned} \tag{3.4}$$

An example of $f_M(m, j)$ for a five-room four-service (i.e., $M = 5$, $N = 4$) system is shown in Table 3.1.

Table 3.1: $f_M(m,j)$ for five patient rooms with four services in each room

$f(m,j)$		m					
		0	1	2	3	4	5
j	2	1	1	1	1	1	1
	3	6	5	4	3	2	1
	4	21	15	10	6	3	1

Therefore, for a system with N services in each patient room, the total number of effective states K will be:

$$K = \sum_{m=0}^{M} f_M(m, N). \tag{3.5}$$

Then, effective states are denoted as $S_l = (n_1^l, n_2^l, \ldots, n_N^l)$, and their steady state probabilities are $P_l = P(n_1^l, n_2^l, \ldots, n_N^l)$, $l = 1, \ldots, K$.

3.3.2 Allocation of Resources to a State

As discussed in Subsection 3.3.1, there are K effective states in the system. For each state S_l, define A_l as the resource availability, $A_l = (a_1^l, a_2^l, \ldots, a_N^l)$, where a_i^l, $i = 1, \ldots, N$, denote the number of resources available for service i for a given state S_l. For example, assume r_1 MAs are available for both the initial service and the wrap-up. In addition, if two of them are working with patients on wrapping-up which has a higher priority than the initial service, then the number of available MAs for the initial service will be $r_1 - 2$. Therefore, a_i^l represents the difference between the quantity of resources for service i and the number of the same resources assigned to a higher priority service, which can be calculated as follows:

$$a_i^l = r_{\theta_i} - \sum_{k=1}^{N} \alpha_{k,i}^l \cdot n_k^l, \tag{3.6}$$

where $\alpha_{k,i}^l$ is an index function of whether there exists another service k having a higher priority than service i for the same resource r_{θ_i}, and $\alpha_{k,i}$ can be calculated from

$$\alpha_{k,i}^l = \begin{cases} 1, & \text{if } \theta_k = \theta_i \text{ and } p_k > p_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3.7}$$

3.3.3 State Transitions

Define β_i^l, $i = 1, \dots, N$, $l = 1 \dots, K$, as the out-going transition rate for service i in state S_l (note that here the self-loop transition rate is ignored). Based on the availability of resources,

$$\beta_i^l = \begin{cases} c_i n_i^l, & \text{if } n_i^l \le a_i^l, \\ c_i a_i^l, & \text{if } n_i^l > a_i^l. \end{cases} \tag{3.8}$$

For a given state S_l, let k_i^l denote the index number corresponding to states $(n_1^l, \dots, n_i^l + 1, n_{i+1}^l - 1, \dots, n_N^l)$, $i = 1, \dots, N-1$, and k_N^l for the state $(n_1^l - 1, n_2^l, \dots, n_N^l + 1)$. The out-going transition probabilities of the current state will be equal to the incoming probabilities from all other states. Then, the following transition equation can be obtained:

$$\begin{aligned} \sum_{i=1}^{N} \beta_i^l P(n_1^l, n_2^l, \dots, n_N^l) &= \sum_{j=1}^{N-1} \beta_j^{k_j^l} P(n_1^l, \dots, n_j^l + 1, n_{j+1}^l - 1, \dots, n_N^l) \\ &+ \beta_N^{k_N^l} P(n_1^l - 1, n_2^l, \dots, n_N^l + 1). \end{aligned} \tag{3.9}$$

Next, introduce a matrix Φ, where for $l = 1, \dots, K$,

$$\Phi(k_i^l, l) = \begin{cases} -\beta_i^{k_i^l}, & \text{if } P_{k_i^l} > 0, \\ 0, & \text{if } P_{k_i^l} = 0, \end{cases} \tag{3.10}$$

and

$$\Phi(l,l) \;=\; \sum_{i=1}^{N} \beta_i^l, \quad i \neq l, \quad l = 1, 2, \cdots, K. \tag{3.11}$$

Thus, a transition matrix Φ with the dimension $K \times K$ and rank $K-1$ is obtained. By taking the first $K-1$ rows of Φ and adding a normalization condition

$$\sum_{l=1}^{K} P_l = 1, \tag{3.12}$$

we construct a new matrix Γ, where

$$\Gamma(l,j) = \Phi(l,j), \quad l = 1, \ldots, K-1, j = 1, \ldots, K, \tag{3.13}$$

$$\Gamma(K,j) = 1, \quad j = 1, \ldots, K. \tag{3.14}$$

Then introduce vectors X and Y, such that

$$X \;=\; [P_1, P_2, \ldots, P_K]^T, \tag{3.15}$$

$$Y \;=\; [0, \ldots, 0, 1]^T. \tag{3.16}$$

We obtain the balance equations as

$$\Gamma X = Y. \tag{3.17}$$

Therefore, the steady state probabilities can be obtained by solving the equation

$$X = \Gamma^{-1} Y. \tag{3.18}$$

Since we consider an irreducible Markov chain with finite number of states, there always exists a unique steady state solution.

3.3.4 Length of Visit and Staff Utilization

To obtain the average patient length of visit, first we evaluate the throughput of the system. Define TP as the rate of patients leaving from the last service, then we have

$$TP = \sum_{l=1}^{K} \beta_N^l P_l, \tag{3.19}$$

where P_l is calculated by solving (3.18).

Since there are M rooms in the system, and each room constantly holds one patient, then, the length of visit can be obtained from Little's Law.

Theorem 3.1 *Under assumptions (i)-(vii), the patient length of visit T_s can be calculated as*

$$T_s = \frac{M}{TP} = \frac{M}{\sum_{l=1}^{K} \beta_N^l P_l}, \tag{3.20}$$

where P_l is an element of vector X, solved from (3.18) and Γ and Y are defined in (3.13)-(3.16).

In addition to the length of visit, the resource utilizations can also be obtained. For a given resource j, assume its quantity is r_j. Let ϵ_i represent that type j resource is requested by service i and η_i indicate that type j resource is assigned to service i. Define ρ_j as the utilization of type j resource, i.e., the long-run percentage of time that type j resource is working. Then we have Theorem 3.2.

Theorem 3.2 *Under assumptions (i)-(vii), the utilization of type j resource can be calculated as*

$$\rho_j = \frac{1}{r_j} \sum_{l=1}^{K} \left(P_l \sum_{i=1}^{N} \epsilon_i \eta_i^l \right). \tag{3.21}$$

where

$$\epsilon_i = \begin{cases} 1, & \textit{if } \theta_i = j, \\ 0, & \textit{otherwise,} \end{cases} \tag{3.22}$$

and

$$\eta_i^l = \begin{cases} 1, & \textit{if } \beta_i^l > 0, \\ 0, & \textit{otherwise.} \end{cases} \tag{3.23}$$

3.4 Discussions

Theorem 3.1 and 3.2 provide a method to evaluate the patient length of visit and staff utilizations in patient rooms, using which the efficiency of care services can be analyzed and the outcome for various values of system parameters can be predicted. In particular, in order to improve the system performance (such as reducing length of visit), understanding the monotonic properties is of importance, which can help us determine which variable to adjust that leads to an improvement of the system performance.

To illustrate the method introduced in Section 3.3, consider the following scenario, where there are three patient rooms, and the care services include nurse's initial check, doctor's visit, and nurse's medication and wrap-up (discharge). This is typical for outpatient clinics and fast track divisions for low acuity patients in EDs. Assume there are two nurses and one doctor, and nurse medication and discharge has a higher priority than the initial check. Then, we have

$$\begin{gathered} M = 3, \qquad N = 3, \qquad R = 2, \\ \mathcal{R} = [2, 1], \quad \Theta = [1, 2, 1], \quad \mathcal{P} = [0.1, 0.2, 0.3]. \end{gathered} \tag{3.24}$$

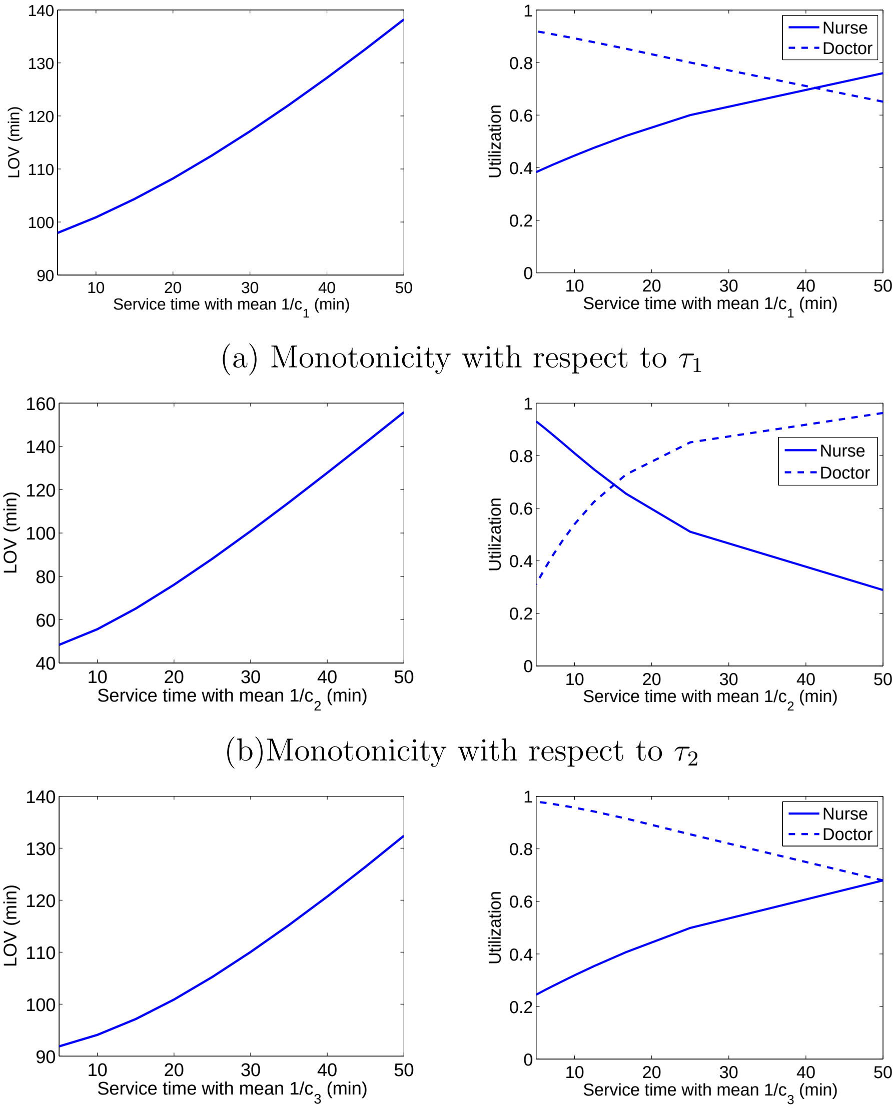
Note that the priority $\mathcal{P}$ can be assigned to any values that ensure $p_3 > p_1$. Next, we study the system with parameters defined in (3.24) as an example to illustrate the monotonic properties.

3.4.1 Monotonicity with respect to Service Time

Assume the means of cycle times τ_1 to τ_3 are increasing from 5 to 50 minutes. As one can see from Figure 3.3, when any service time is increased, the patient LOV is increasing as well. A shorter service time leads to a reduced flow time of patients, which agrees with our intuition. Similarly, the monotonicity of care provider's utilization with respect to service time exists as well. With the increase of service time of one provider (such as nurse), her/his utilization is increasing, while the utilization of the other one (respectively, doctor) is decreasing, since less throughput is obtained.

3.4.2 Monotonicity with respect to Care Provider Quantity

Next we consider the monotonicity with respect to resource quantity. Alter the numbers of nurses and doctors and change from one to four, respectively. Note that for the sake of completeness, we allow the number of providers to go beyond the room number. As illustrated in Figure 3.4, when the numbers of care providers increase, the patient LOV decreases. Specifically, when the number of nurses is increasing from one to two, there is a dramatic decrease in LOV. When the nurse quantity is increased to three, such drop is modest. Further increase of nurse quantity does not lead to any change in LOV since now each room is already assigned one nurse. Similar observations are obtained when the number of doctors is increased. When the quantities of staff exceed the number of patient rooms, the LOV will no longer be reduced. In addition, increasing the quantity of one provider (e.g., nurse), their utilization is decreased, while the utilization of the



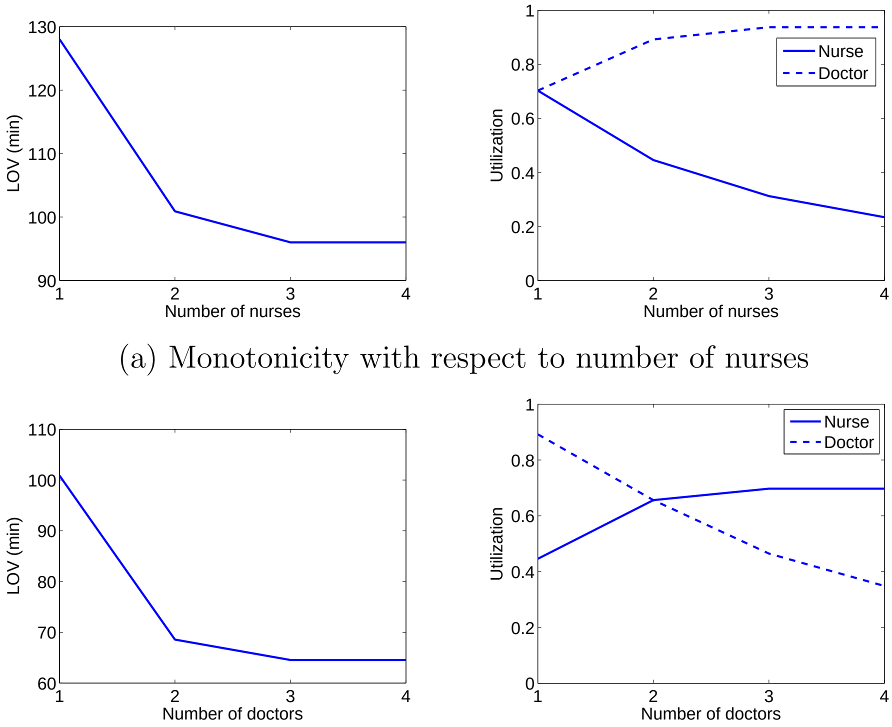
(a) Monotonicity with respect to τ_1

(b)Monotonicity with respect to τ_2

(c) Monotonicity with respect to τ_3

Figure 3.3: Monotonicity of LOV and utilization w.r.t. service times

other provider (respectively, doctor) is increasing due to a higher throughput. However, there is a diminishing return of the increasing utilization, because when the number of nurses (respectively, doctors) reaches three, the maximum throughput has been achieved (since each room only allows one patient). Therefore, the utilization (of doctor) will not increase anymore.



(a) Monotonicity with respect to number of nurses

(b) Monotonicity with respect to number of doctors

Figure 3.4: Monotonicity of LOV and utilization w.r.t. number of care providers

3.4.3 Monotonicity with respect to Patient Room Quantity

Now we investigate the monotonicity as a function of number of patient rooms. As shown in Figure 3.5, increasing the number of patient rooms from three to ten, the

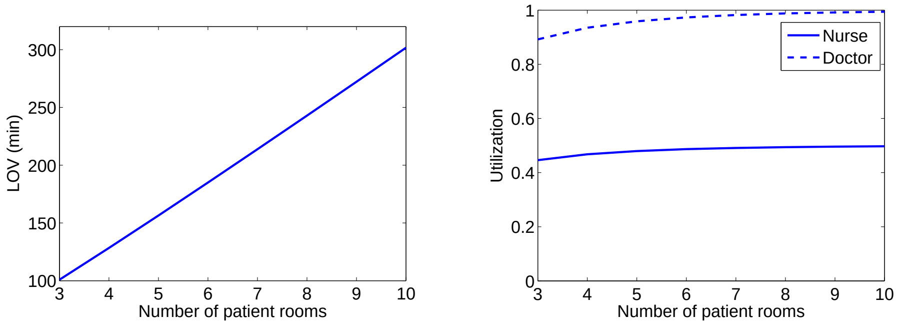


Figure 3.5: Monotonicity of LOV and utilization w.r.t. number of patient rooms

patient LOV exhibits an almost linearly increasing trend. In addition, the doctor's utilization is increased to almost 100% when more patient rooms are added. However, the nurse's utilization keeps flat since now the doctor is the bottleneck of the system, which limits the patient throughput.

3.4.4 Two-Room Two-Service Case

In the case of two patient rooms with each having two services, some of the system-theoretic properties can be proved analytically.

Closed-form expressions

For the two-room two-service case, closed-form expressions for length of visit and staff utilizations are available. Such cases represent the scenarios that low-acuity patients receive express services in EDs and outpatient clinics (e.g., only nurse visit and doctor diagnosis are needed). In this case, we obtain Corollary 3.1.

Corollary 3.1 *Under assumption (i)-(vii) with $M = 2$, $N = 2$ and $r_1 = r_2 = 1$, the* *patient length of visit T_s and resource utilization ρ_i can be calculated as*

$$\begin{aligned} T_s &= \frac{2(\tau_1^2+\tau_2^2+\tau_1\tau_2)}{\tau_1+\tau_2}, & (3.25) \\ \rho_1 &= \frac{\tau_1(\tau_1+\tau_2)}{\tau_1^2+\tau_2^2+\tau_1\tau_2}, & (3.26) \\ \rho_2 &= \frac{\tau_2(\tau_1+\tau_2)}{\tau_1^2+\tau_2^2+\tau_1\tau_2}. & (3.27) \end{aligned}$$

Proof: See the Appendix. ■

Monotonic properties

Corollary 3.2 *Under assumptions (i)-(vii) with $M = 2$, $N = 2$ and $r_1 = r_2 = 1$, the patient length of visit is monotonically increasing with respect to service time, i.e., T_s is strictly increasing with respect to τ_1 or τ_2.*

Proof: See the Appendix. ■

Corollary 3.3 *Under assumptions (i)-(vii) with $M = 2$, $N = 2$ and $r_1 = r_2 = 1$, the staff utilization is monotonically increasing with respect to its own service time, but decreasing with the other service time, i.e., ρ_i is strictly increasing with respect to τ_i, but strictly decreasing with respect to τ_j, $j \neq i$.*

Proof: See the Appendix. ■

Corollary 3.2 and 3.3 are consistent with the results obtained in Subsection 3.4.1.

Bottleneck service

A service is referred to as the most impeding service (or bottleneck service) if its improvement can lead to the most significant improvement in system performance (e.g., length of visit). To determine which service may have the largest impact on patient length of visit, compare $\frac{\partial T_s}{\partial \tau_1}$ and $\frac{\partial T_s}{\partial \tau_2}$. Then the following results are obtained:

Corollary 3.4 *Under assumptions (i)-(vii) with $M = 2$, $N = 2$ and $r_1 = r_2 = 1$, the service with the longer service time is the more impeding service (i.e., bottleneck).*

Proof: See the Appendix. ■

Identifying bottlenecks and eliminating them are the most effective way to improve the operational efficiency in manufacturing systems ([130]). In healthcare systems, such an approach has also demonstrated its effectiveness to improve patient flow and reduce patient length of visit (see, for instance, [17], [19]). The results obtained here for the two-room two-service case also match our intuition. However, for more complicated cases, i.e., $M > 2$, $N > 2$, and $r_1 \neq r_2$, direct comparison among service times may not lead to the precise identification of bottleneck services. Therefore, developing an appropriate and effective method to identify the service bottlenecks is of importance and will be investigated in future work.

Service allocation

In many cases, it is of interest to investigate the principle for assigning workload to nurses and doctors more efficiently. If with well trained nurses, some of the services can be assigned to either a nurse or a doctor, then, the most efficient way to minimize patient length of visit is to balance the work between the doctor and nurse (if it is possible). In other words, given the constraint that the total service time is a constant, we obtain:

Corollary 3.5 *Under assumptions (i)-(vii) with $M = 2$, $N = 2$ and $r_1 = r_2 = 1$, under the constraint of $\tau_1 + \tau_2 =$ constant, the optimal allocation of services is $\tau_1^* = \tau_2^*$.*

Proof: See the Appendix. ■

Notably, workload assignment will be constrained by the nature of work, staff capabilities, and other factors. However, the results obtained here indicate that an effort to balance the workload will be beneficial in reducing patient length of visit. More in-depth analysis is desired in future work for larger systems.

Joint service

The method introduced in this chapter can be extended to model the scenario of joint service, where more than one providers are needed to carry out the service. For instance, the doctor may need the nurse's help to carry out certain treatments, or the resident doctor may need the supervision of the attending physician. In these scenarios, the states of the system need modification – number of states is reduced due to joint service. As an illustration of this, consider a two-room two-service (e.g., nurse visit, and nurse/doctor joint service) system, with one nurse and one doctor. Assume that the joint service has a higher priority than the nurse's initial visit. In this case, it is easy to show that:

Corollary 3.6 *Under assumptions (i)-(vii) with $M = 2$, $N = 2$ and $r_1 = r_2 = 1$, and assume the second service needs both providers, then the patient length of visit T_s and* *resource utilization ρ_i can be calculated as*

$$T_s = 2(\tau_1 + \tau_2), \tag{3.28}$$

$$\rho_1 = 100\%, \tag{3.29}$$

$$\rho_2 = \frac{\tau_2}{\tau_1 + \tau_2}. \tag{3.30}$$

Proof: See the Appendix. ■

Comparing with Corollary 3.1, the length of visit is longer due to the fact that only one patient can be served at any time. In other words, the first provider (nurse) is always busy serving patients, and the second provider (doctor) will be idle when the nurse is rooming the patient. The monotonicity still holds, i.e., T_s is monotonically increasing with respect to τ_1 and τ_2, and ρ_2 is monotonically increasing with respect to τ_2, and decreasing with respect to τ_1.

3.5 Extensions to Non-Exponential Scenarios

The analysis presented in Sections 3.3 and 3.4 assumes exponential service times. In practice, assumptions of this type may not hold. Therefore, in this section, we extend the study to incorporate the scenarios of non-exponential service times.

3.5.1 Dependency on Distribution Type

To study the case of the non-exponential service time, the following two questions need to be answered. First, is the system performance dependent on the distribution type of the service time or not? If it is dependent, then it implies that a formula or an approach is needed for each distribution type, which makes the analysis intractable. If it is the

Table 3.2: Types of mixed distributions

Mix 1	Mix 2	Mix 3	Mix 4
Uni, Ln, ga	Ln, Uni, Ga	Ln, Ga, Uni	Ga, Tri, Uni

(a) System 1

Mix 5	Mix 6	Mix 7	Mix 8
Ln, Ga, Tri	Ga, Tri, Uni	Ga, Uni, Tri	Tri, Uni, Ln

(b) System 2

opposite, then whether or not there exists an empirical or approximation formula for performance evaluation?

To answer these two questions, extensive simulation experiments have been carried out. Dozens of examples have been selected randomly to compare the lengths of visit under several commonly used service time distributions such as uniform (Uni), triangular (Tri), lognormal (Ln), and gamma (Ga) distributions, and a mix of them. Identical mean and coefficient of variation (CV) are assumed for each distribution.

In all the examples we have tested, the differences in length of visit are very limited (typically within 2%). This indicates that, practically, the length of visit is independent of higher distribution moments (such as the third distribution moment and above), but mainly depends on the mean and CV. Table 3.3 illustrates this property observed from two four-room three-service systems with service times under uniform, triangular, lognormal, gamma, and mixed distributions (denoted as System 1 and System 2, and $CV = 0.4$ and 0.55 in Systems 1 and 2, respectively). The mixed distributions are summarized in Table 3.2, meaning that the probability distribution for each of the three services is randomly selected from the aforementioned distributions.

Table 3.3: Impact of distribution type

Distribution i	Uni	Tri	Ln	Ga
$T_{s,i}$	364.2	364.2	362.4	362.6

Distribution i	Mix 1	Mix 2	Mix 3	Mix 4
$T_{s,i}$	360.3	362.7	364.25	363.7

(a) System 1

Distribution i	Uni	Tri	Ln	Ga
$T_{s,i}$	373.3	373.3	367	371

Distribution i	Mix 5	Mix 6	Mix 7	Mix 8
$T_{s,i}$	372.75	372.3	372.2	370.26

(b) System 2

Define δ to quantify the maximum relative difference among lengths of visit $T_{s,i}$ under distribution i, $i \in \{\text{Uni, Tri, Ln, Ga, Mix 1, \ldots, Mix 4}\}$,

$$\delta = \frac{\max_i T_{s,i} - \min_i T_{s,i}}{\frac{1}{8}\sum_i T_{s,i}} \cdot 100\%.$$

For Systems 1 and 2, $\delta = 1.1\%$ and 1.7%, which are negligible.

The above results indicate that the patient length of visit is practically independent of the higher distribution moments of the service time, but mainly dependent on the mean and coefficient of variation of the service time. This result is also consistent with the conclusion obtained by Reynolds et al. [18]. Note that here we focus on $0 < CV < 1$, since for most services, the longer time the service has been carried out, the more likely the service will be finished, which leads to $CV < 1$ [131].

3.5.2 Empirical Formula

To evaluate the length of visit, an empirical formula is pursued. First, define an effective CV^2 which is a weighted average of all service CV squares:

$$CV_{\text{eff}}^2 = \frac{\sum_{i=1}^{M} \tau_i cv_i^2}{\sum_{i=1}^{M} \tau_i}. \tag{3.31}$$

In Table 3.4, the first two columns present the CV_{eff}^2's and the corresponding estimates of the length of visit by simulating two empirical systems (denoted as System 3 and System 4). As one can see, the length of visit increases with CV_{eff}^2 approximately linearly. This is also observed in all the empirical studies we've conducted. Therefore, it can be hypothesized that the length of visit for any CVs between 0 and 1, $T_s^{non-exp}$, can be evaluated through the following empirical formula:

$$T_s^{non-exp} = T_s^0 + (T_s^{exp} - T_s^0)CV_{\text{eff}}^2, \tag{3.32}$$

where T_s^{exp} can be calculated using Theorem 3.1, and

$$T_s^0 = \max\left\{ M \cdot \max_{j=1,\ldots,R} \left(\sum_{\substack{i=1,\ldots,N \\ s.t.\ \theta_i = j}} \frac{\tau_i}{r_j} \right), \sum_{i=1}^{N} \tau_i \right\}, \tag{3.33}$$

where M, N, R are the number of rooms, number of services in each room, and provider types, respectively; r_i is the number of type i providers, and θ_i defines the type of provider assigned to service i. Basically, T_s^0 intends to approximate the length of visit when there is no variability in service time in an ideal case, i.e., the utilization of a provider is maximized. Then T_s^0 is determined by the total time for a patient going through all the services, or the time needed by one provider type to visit all the patient rooms, whichever is longer.

Based on extensive simulation experiments, we have observed that the empirical formula (3.32) provides an accurate estimate of the length of visit in most practical cases. In all the examples we randomly generated, the maximum relative difference between the simulation and analytical models is less than 2%. In Table 3.4, the third column presents the lengths of visit obtained by the empirical formula, and the fourth column illustrates the accuracy of using the empirical formula (3.32) for evaluating the two systems (System 3 and System 4) compared with simulation. ϵ is the relative difference, defined as

$$\epsilon = \frac{T_s^{empirical} - T_s^{sim}}{T_s^{sim}} \cdot 100\%. \tag{3.34}$$

In Table 3.4, all ϵ's are very small. It can be concluded that the empirical formulas (3.32) provide accurate estimates of patient length of visit within patient rooms. In addition, the utilization of care providers can be evaluated based on the time percentage of services for a provider in a given time period. It can be derived from the average

Table 3.4: Accuracy of the empirical formula

CV^2_{eff}	T_s^{sim}	$T_s^{empirical}$	ϵ
0.0025	360	360.1	0.03%
0.0225	359.9	361.2	0.33%
0.0625	359.9	363.2	0.92%
0.1225	361.2	366.3	1.42%
0.2025	365.4	370.4	1.39%
0.3025	371	375.6	1.25%
0.4225	379.2	381.8	0.71%
0.5625	387.7	389.1	0.35%
0.7225	397.2	397.4	0.06%
0.9025	406.7	406.7	0.01%

(a) System 3

CV^2_{eff}	T_s^{sim}	$T_s^{empirical}$	ϵ
0.0025	390	390.2	0.06%
0.0225	390	391.7	0.46%
0.0625	391.3	394.8	0.91%
0.1225	396.6	399.4	0.74%
0.2025	404.6	405.6	0.26%
0.3025	415.2	413.4	-0.43%
0.4225	428	422.7	-1.24%
0.5625	439.6	433.5	-1.39%
0.7225	453.3	445.8	-1.63%
0.9025	465.5	459.8	-1.22%

(b) System 4

number of patients in the period (calculated from $T_s^{non-exp}$) and the average service time for each patient, i.e.,

$$\rho_i = \frac{M}{T_s^{non-exp}} \sum_{\substack{j = 1, \ldots, N \\ s.t. \theta_j = i}} \tau_j. \tag{3.35}$$

3.6 Conclusions

In this chapter, an analytical framework is introduced to model care delivery operations within patient rooms. Critical performance measurements such as patient average length of visit and staff utilizations can be evaluated. The extension to non-Markovian scenarios makes the model applicable for more generalized systems. Using the proposed model, health care professionals can evaluate different design options corresponding to capacity planning, workforce configuration, and service delivery, etc.

Chapter 4

Modeling of Care Delivery Operations with Shared Resources

4.1 Introduction

The analytical model introduced in Chapter 3 provides a viable method for evaluating care delivery operations within patient rooms. However, like all other Markovian models, such an approach may experience the curse of dimensionality when multiple non-identical rooms, and in particular, with shared resources, are considered. Therefore, in this chapter, we introduce a system-theoretic approach based on the Markov chain analysis to model the care delivery system with shared resources. The care delivery system featuring multiple exam rooms and limited resources is described in Section 4.2. A Markov chain model of the patient flow in a single exam room is developed using the same modeling framework as described in Chapter 3. In Section 4.3, a converging iterative method, referred to as the shared resource iteration for evaluating the scenario of two or more non-identical exam rooms is introduced. The convergence of the recursive procedure is justified. The approximation formulas for evaluating the non-Markovian scenarios are sketched.

4.2 System Description and Modeling

4.2.1 Process Description

Care delivery systems are highly stochastic and interdependent. They usually consist of multiple exam rooms and complex staff configurations, and provide a variety of care services. Typically, care is offered by a team which comprises of physicians, nurses, and medical assistants, or technologists and technologist assistants in diagnostic imaging and laboratory test centers. The capacity of these resources varies according to practice and facility sizes, demands, and purposes. Besides, it is often the case that supportive staff are shared among chief care providers. For example, in primary care clinics, medical assistants can be shared by several physicians within one pod of the clinic. Similarly, during a testing procedure, technologist assistants take care of patients belonging to different technologists.

As described in Chapter 3, within each exam room, only one patient is permitted at a time and the care services are carried out sequentially. However, although each chief care provider works independently, the supportive staff are shared among all the exam rooms, which may introduce an availability issue and cause additional delay. To model the process described in this subsection, the notations, assumptions and problem formulation are introduced in Subsection 4.2.2.

4.2.2 Notations, Assumptions and Problem Formulation

In this subsection, we consider a care provider team consists of several chief care providers and one supportive staff. The following notations are introduced to address the services, resources, and their interactions.

- The workflow of the care delivery process follows steps 1 to 4 sequentially:
  - (1) patient checking in,
  - (2) supportive staff rooming the patient,
  - (3) chief care provider serving the patient (diagnose or conduct test),
  - (4) patient checking out.
- The number of exam rooms in the system is denoted as M. Each room can only accommodate one patient.
- A patient needs to wait in the lobby (waiting room) if he/she arrives early and is blocked by previous patients, or the supportive staff is not available. The maximum capacity of waiting for each exam room is denoted as Q_i, $i = 1, 2, \ldots, M$. Without loss of generality, we start with $Q_i = 5$ and extend to larger numbers as appropriate.
- The number of chief care providers is defined as r_1, and the number of supportive staff is denoted as r_2.

In addition, the following assumptions regarding the arrival and services are introduced:

1) The inter-arrival time of the incoming patients for each room follows the exponential distribution with arrival rate λ_i, $i = 1, 2, \ldots, M$.

2) The four processes each patient has to complete, i.e., patient check-in, supportive staff rooming, chief care provider diagnosis and patient check-out, are denoted as services 1, 2, 3 and 4, respectively. The service time for service j in exam room i is exponentially distributed with mean processing time $\tau_{j,i}$, $j = 1, \ldots, 4$, $i = 1, 2, \ldots, M$.

Then, the corresponding processing rate is $c_{j,i} = \frac{1}{\tau_{j,i}}$. Notably, non-identical service times are assumed for different patient rooms. The identical service time is a special case of this assumption.

3) The resource cannot be released until the service in progress finishes. An ongoing service cannot be disrupted before completion, i.e., if the resource is busy, the next patient has to wait until the current service is completed.

4) In the current system, each chief care provider is dedicated to one exam room, and the supportive staff is shared for all rooms, and thus, $r_1 = M$ and $r_2 = 1$.

Instead of modeling the system directly, we start with each room individually, denoted as the subsystem. In an appropriately defined state space, the subsystem satisfying assumptions 1)-4) is a stationary random process. Let T_i , $i = 1, 2, \ldots, M$, denote the patient length of visit for each exam room in the subsystem. In the framework of 1)-4), T_i is a function of all system parameters:

$$T_i = f_T(Q_i, c_{1,i}, c_{2,i}, c_{3,i}, c_{4,i}, \lambda_i). \tag{4.1}$$

Denote the staff utilizations as $\rho_{1,i}$ and $\rho_{2,i}$, for the care provider and supportive staff in room i, respectively, which are also functions of all system parameters:

$$\rho_{r,i} = f_\rho(Q_i, c_{1,i}, c_{2,i}, c_{3,i}, c_{4,i}, \lambda_i),\ r = 1, 2. \tag{4.2}$$

Then, the problem to be addressed is: *Under assumptions 1)-4), develop a method to evaluate the patient length of visit T_i, and staff utilizations $\rho_{1,i}$ and $\rho_{2,i}$ as functions of system parameters.*

Solutions to the problem are presented in Section 4.2.3.

4.2.3 Performance Analysis

Let $S = \{s_1, s_2, s_3, s_4\}$ denote the states of the subsystem, where s_i represents the number of patients in stage i, $i = 1, \ldots, 4$. For example, $s_2 = m$ indicates that there are m patients in process 2 (supportive staff rooming), either being served by the supportive staff or waiting to be served by the chief care provider. The feasible states satisfy the following constraints:

- $s_j \geq 0$, $j = 1, \ldots, 4$,
- $s_1 \leq Q_1$, queue length constraint,
- $s_2 + s_3 \leq 2$, resource constraint,
- $s_4 \leq 1$, resource constraint.

The total number of feasible states is denoted as K. It is a function of the queue length and the possible allocations of patients in the subsystem. When the queue length is zero, we have twelve states: (0,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1), (0,1,1,0), (0,0,1,1), (0,1,0,1), (0,0,2,0), (0,2,0,0), (0,0,2,1), (0,2,0,1), (0,1,1,1). If the queue length is Q_1, then we have $(Q_1 + 1) \times 12$ states. In the subsystem model, $Q_1 = 5$, so there are 72 feasible states. Then the steady state probability for a feasible state S_k, $k = 1, \ldots, K$, is defined as

$$P_k = P(s_1^k, s_2^k, s_3^k, s_4^k), \quad k = 1, 2, \ldots, K.$$

The state transition can be triggered by one of the following events: (1) patient arrival; (2) patient check-in; (3) supportive staff finishing rooming; (4) care service completion; and (5) patient check-out. Note that these events cannot occur simultaneously. To illustrate the transition, for a state $S_k = \{s_1^k, s_2^k, s_3^k, s_4^k\}$, the following events can occur.

- When a patient arrives, if the waiting room is full (although unlikely), i.e., $s_1^k = Q$, then the patient is lost due to space limit, and no transition will occur. Otherwise, the current state goes to state $\{s_1^k + 1, s_2^k, s_3^k, s_4^k\}$ with arrival rate λ.

- When a patient checks in, if the exam room is occupied or there is a patient waiting for rooming already, i.e., $s_2^k + s_3^k = 2$, then no transition will occur and the patient remains waiting in the lobby. Otherwise, the patient goes to a sub-waiting room after checking-in, and the state S_k has a rate c_1 to transit to state $\{s_1^k - 1, s_2^k + 1, s_3^k, s_4^k\}$. Note that such an event occurs only when there is a patient ready to check in, i.e., $s_1^k \geq 1$.

- With the completion of rooming, the patient leaves the sub-waiting room and either starts or waits for diagnosis with rate c_2. Thus, the state S_k transits to state $\{s_1^k, s_2^k - 1, s_3^k + 1, s_4^k\}$ with rate c_2 when $s_2^k \geq 1$.

- After all care services being delivered, if $s_4^k = 1$, then the patient waits for checking out and no transition occurs. Otherwise, he/she goes to check out and S_k transits to state $\{s_1^k, s_2^k, s_3^k - 1, s_4^k + 1\}$ with rate c_3 and $s_3^k \geq 1$.

- Finally, when the patient checks out, S_k moves to state $\{s_1^k, s_2^k, s_3^k, s_4^k - 1\}$ with rate c_4. Again, such an event occurs only when there is a patient ready to check out, i.e., $s_4^k \geq 1$.

Let $\Phi(k, l)$ define the transition rate from state S_k to S_l, $k, l = 1, \ldots, K$, $k \neq l$, which takes one of the values of c_1, c_2, c_3, c_4, or λ. In addition,

$$\Phi(l, l) = -\sum_{j=1}^{K} \Phi(l, j), \quad j \neq l, \quad l = 1, 2, \ldots, K. \tag{4.3}$$

Thus, a transition matrix Φ with the dimension $K \times K$ and rank $K - 1$ is obtained. Similar to the derivation described in Section 3.3.3, the steady state probability can be calculated.

Define TP_i as the throughput rate of subsystem i, i.e., the rate patient leaving from the last service in room i, and define WIP_i as the average number of patients in the subsystem. According to Little's Law, the average patient length of visit in room i, T_i, can be obtained and the evaluation formulas are articulated in Theorem 4.1.

Theorem 4.1 *Under assumptions 1)-4) with exam room i,*

$$T_i = \frac{WIP_i}{TP_i} = \frac{\sum_{l=1}^{K}\left(P_{l,i}\sum_{j=1}^{4} s_{j,i}^{l}\right)}{c_{4,i}\sum_{l=1}^{K} P_{l,i}s_{4,i}^{l}}, \tag{4.4}$$

where $P_{l,i}$ is solved using the same approach introduced in Section 3.3.3 and

$$TP_i = c_{4,i}\sum_{l=1}^{K} P_{l,i}s_{4,i}^{l}, \tag{4.5}$$

$$WIP_i = \sum_{l=1}^{K}\left(P_{l,i}\sum_{j=1}^{4} s_{j,i}^{l}\right). \tag{4.6}$$

In addition to the patient length of visit, the staff utilizations can be calculated as

Theorem 4.2 *Under assumptions 1)-4) with exam room i,*

$$\rho_{1,i} = \sum_{l=1}^{K} P_{l,i}\cdot I_{\{s_{3,i}^{l}>0\}}, \tag{4.7}$$

$$\rho_{2,i} = \sum_{l=1}^{K} P_{l,i}\cdot I_{\{s_{2,i}^{l}>0\}}. \tag{4.8}$$

Remark 4.1 The utilization of the supportive staff and the chief care provider only captures the time percentage they spend directly encountering patients. Typically they have multiple job duties, which are not included in the current model.

4.3 Shared Resource Iteration

4.3.1 Two Non-Identical Rooms

The results obtained in Subsection 4.2.3 only consider one exam room by assuming that the supportive staff is always available. However, when the system consists of two or more exam rooms, which are not necessarily identical, the supportive staff will be shared by all exam rooms. In the case of two exam rooms, a sample workflow is illustrated in Figure 4.1, where the rectangles represent the processes and the dash rectangle indicates that the rooming processes share the same resource.

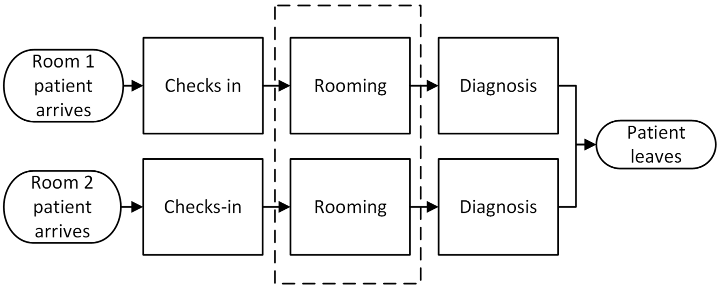


Figure 4.1: Patient flow model: two exam rooms

Remark 4.2 Note that although all the patients check in at the same reception desk, the registration time is short and the receptionists are typically available when an individual patient arrives, so we do not view it as a process with resource constraint.

By just adding one exam room, the complexity of the system increases enormously. The size of the state space increases significantly and the transitions are subject to resource constraints. Therefore, to evade the efforts of expanding the Markov chain model and provide the possibility for the further extension to more exam rooms, an iterative approach, nominated as the shared resource iteration, is introduced. First, let p_i, $i = 1, 2$, be the time percentage the supportive staff spends in room i during a patient's visit. Assume the initial time percentage is known, denoted as $p_1^{(0)}$, $p_2^{(0)}$, where the subscript indicates the room number, and the superscript represents the iteration number (0 refers to the initial value). By cause of sharing, the transition rate of the supportive staff's service in room 1 is decreased to accommodate the extra waiting due to supportive staff's service in room 2. Thus, in the first iteration, we have

$$c_{2,1}^{(1)} = c_{2,1}(1 - p_2^{(0)}).$$

Then the average length of visit for a patient in room 1 can be calculated using Proposition 4.1, denoted as function $f_T(\cdot)$.

$$T_1^{(1)} = f_T(Q_1, c_{1,1}, c_{2,1}^{(1)}, c_{3,1}, c_{4,1}, \lambda_1).$$

Also obtained is the time percentage the supportive staff spends in room 1,

$$p_1^{(1)} = \frac{1}{c_{2,1} T_1^{(1)}}.$$

Using this new time percentage, we analyze another subsystem room 2. The transition rate for room 2 is updated as

$$c_{2,2}^{(1)} = c_{2,2}(1 - p_1^{(1)}).$$

Similarly, the length of visit and supportive staff's time percentage in room 2 can be calculated:

$$\begin{aligned} T_2^{(1)} &= f_T(Q_2, c_{1,2}, c_{2,2}^{(1)}, c_{3,2}, c_{4,2}, \lambda_2), \\ p_2^{(1)} &= \frac{1}{c_{2,2} T_2^{(1)}}. \end{aligned}$$

This finishes the first iteration. In the next iteration, using the updated probabilities $p_i^{(1)}$, $i = 1, 2$, we repeat the procedure to obtain a new set of estimates $p_i^{(2)}$, and continue.

The iteration terminates when all the differences between p_i^k and p_i^{k+1} are sufficiently small. Then, the corresponding $T_1^{(k)}$ and $T_2^{(k)}$ are the estimated average lengths of visit for the two exam rooms, obtained through Proposition 4.1. Such a process can be represented by the procedure below:

Procedure 4.1

$$\begin{aligned}
c_{2,1}^{(k)} &= c_{2,1}(1 - p_2^{(k-1)}), \\
T_1^{(k)} &= f_T(Q_1, c_{1,1}, c_{2,1}^{(k)}, c_{3,1}, c_{4,1}, \lambda_1), \\
p_1^{(k)} &= \frac{1}{c_{2,1}T_1^{(k)}}, \qquad (4.9) \\
c_{2,2}^{(k)} &= c_{2,2}(1 - p_1^{(k)}), \\
T_2^{(k)} &= f_T(Q_2, c_{1,2}, c_{2,2}^{(k)}, c_{3,2}, c_{4,2}, \lambda_2), \\
p_2^{(k)} &= \frac{1}{c_{2,2}T_2^{(k)}}, \qquad (4.10) \\
k &= 1, 2, \ldots,
\end{aligned}$$

with the initial condition

$$p_2^{(0)} = 0.$$

The convergence of the procedure has been investigated.

Proposition 4.1 *Under assumptions 1)-4), Procedure 4.1 is convergent and*

$$T_i^{exp} = \lim_{k\to\infty} T_i^{(k)}, \quad i = 1, 2. \qquad (4.11)$$

where T_i^{exp} represents the patient length of visit at room i under exponential assumptions of service and inter-arrival times, denoted by the superscript 'exp'.

Proof: See the Appendix. ■

Note that the initial condition $p_2^{(0)} = 0$ is introduced for the proof of convergence. By randomly selecting values of $p_2^{(0)} \in (0, 1)$ during tests, it is observed that the procedure always leads to the same convergent value. In addition, it usually takes three to five iterations for Procedure 4.1 to converge.

4.3.2 Extensions to Multiple Non-Identical Rooms

Next, we evaluate more general scenarios where systems have more than two exam rooms. The workflow is illustrated in Figure 4.2. Directly developing a Markov chain model of the entire system is impossible, and may face a state space explosion when the number of subsystems is large or the interaction among subsystems is complex. Therefore, an extension of the iterative method is desired.

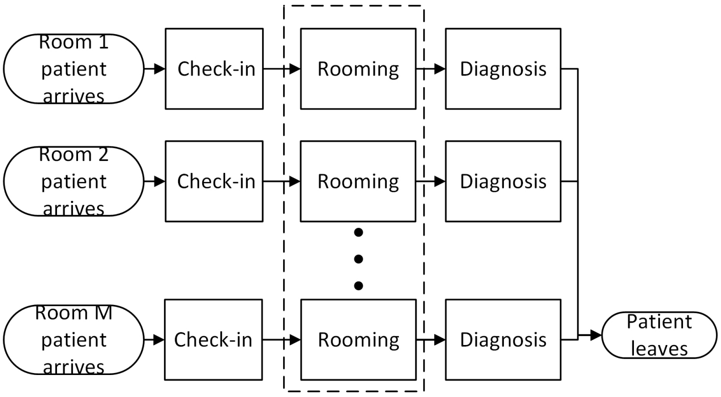


Figure 4.2: Patient flow model: multiple exam rooms

The recursive procedure for more than two exam rooms is similar to that of the system with two exam rooms. The main difference is that when the transition rate is updated for each exam room, the cases that the supportive staff is serving in any exam room need to be considered. Such a modification iterates among all exam rooms. The procedure is presented as follows:

Procedure 4.2

$$\begin{aligned}
c_{2,i}^{(k)} &= c_{2,i}\Bigg(1-\sum_{j=1}^{i-1}p_j^{(k)}-\sum_{j=i+1}^{M}p_j^{(k-1)}\Bigg),\\
T_i^{(k)} &= f_T(Q_i,c_{1,i},c_{2,i}^{(k)},c_{3,i},c_{4,i},\lambda_i),\\
p_i^{(k)} &= \frac{1}{c_{2,i}T_i^{(k)}},\\
i &= 1,2,\ldots,M\\
k &= 1,2,\ldots,
\end{aligned} \tag{4.12}$$

with the initial conditions

$$p_i^{(0)}\in[0,1),\quad \textit{and}\quad \sum_{i=1}^{M}p_i^{(0)}<1,\quad i=1,2,\ldots,M.$$

The convergence property of such systems still holds and is verified through extensive numerical tests using empirical data. We randomly select the number of exam rooms M, service rates $c_{j,i}$ (1/min), and arrival rates λ_i (1/min) from the following sets which represent the typical range of the parameters.

$$\begin{aligned}
M &\in \{2,3,4,5\},\\
c_{j,i} &\in (0.025,0.1),\quad j=1,2,3,4, i=1,\ldots,M,\\
\lambda_i &\in (0.01,0.03),\quad i=1,\ldots,M.
\end{aligned} \tag{4.13}$$

The convergence criterion is met when the differences in p_i between two iterations are small.

$$|p_i^{(k)}-p_i^{(k+1)}|<10^{-5},\quad i=1,\ldots,M.$$

In all the cases we have tested, the convergence of p_i and T_i is observed. Thus, we conclude that Procedure 4.2 is convergent and the following limits exist:

$$T_i^{exp}=\lim_{k\to\infty}T_i^{(k)},\quad i=1,2,\ldots,M. \tag{4.14}$$

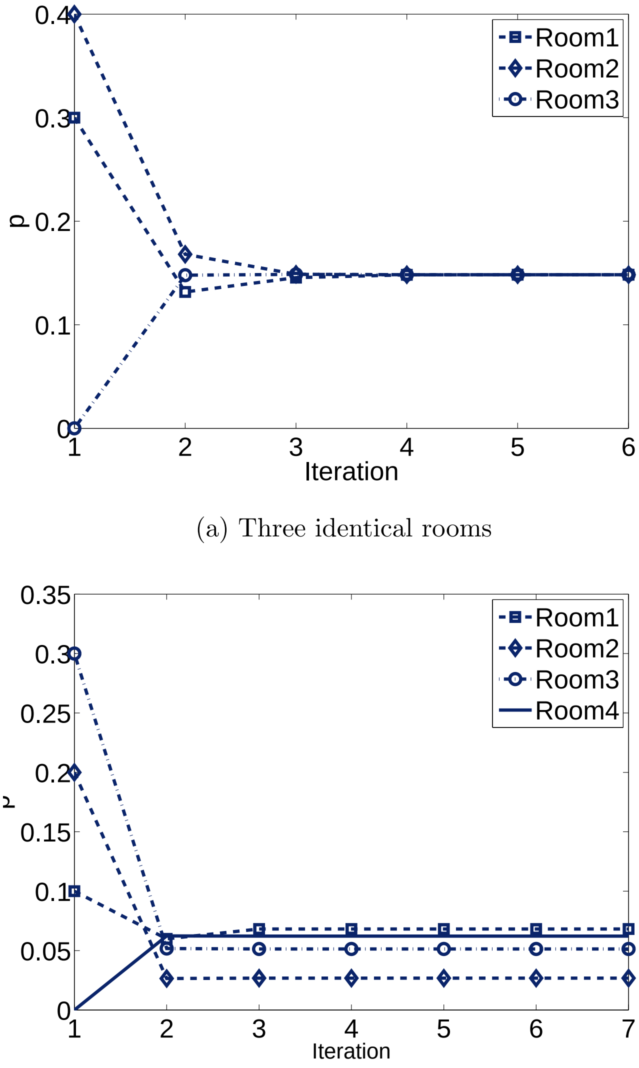
To illustrate the convergence property, the iterations of p_i (i.e., the inverse of T_i divided by the service rate) are shown in Figure 4.3, for three identical rooms and four non-identical rooms scenarios. In Figure 4.3 (a), the time percentages the supportive staff spent in rooms 1-3 are denoted as broken lines with square, diamond, and circle, respectively. The horizontal axis is slotted by iterations. Thus, the dynamic of p_i during each iteration is exhibited in the figure. As one can see, all p_i's converge after roughly three iterations and they all converge to the same value since the three rooms are identical. In Figure 4.3 (b), broken lines with square, diamond, and circle represent p_1 to p_3 in rooms 1-3, respectively, and the solid line illustrates p_4. Again all p_i's converge after about three iterations. However, since all rooms are not identical, these p_i's are not the same. In summary, the convergence of p_i can be observed only after several iterations.

Remark 4.3 This iterative method could be conveniently extended to the system with multiple number of shared resources scenario. For instance, it can be the case that two tech assistants are shared by three technologists. To simplify the analysis, assume each shared resource (e.g., supportive staff) is identical and independent. Similar to the single shared resource case, a transition can fail due to that all the resources are working with other patients. Then the probability that all resources are not available can be represented by $(q^{(k)})^s$, where s is the number of shared resources (such as the number of supportive staff), and $q^{(k)}$ denotes the probability that the resource is not available as the only shared resource, during the k-th iteration. For patient room i,

$$q_i^{(k)} = \sum_{j=1}^{i-1} p_j^{(k)} + \sum_{j=i+1}^{M} p_j^{(k-1)}. \tag{4.15}$$

Then the processing rate $c_{2,i}$ in Procedure 4.2 should be updated as

$$c_{2,i}^{(k)} = c_{2,i}(1 - (q_i^{(k)})^s). \tag{4.16}$$



(a) Three identical rooms

(b) Four non-identical rooms

Figure 4.3: Convergence of p_i

The other formulas in Procedure 4.2 will remain the same. To investigate the accuracy of this iterative method, extensive experiments have been conducted. It has been demonstrated that the performance of this algorithm is similar to that of a single resource for sharing.

4.3.3 Extensions to Non-Exponential Scenarios

The analysis in Subsections 4.3.1 and 4.3.2 assumes exponential inter-arrival and service times. However, the exponential assumption may not hold in practice. Similar to what has been discussed in Section 3.5, the modeling framework developed in this chapter is also extended to non-exponential scenarios. The derivations are skipped and only the empirical formulas are provided.

Empirical formula

If the scheduled inter-arrival time is long enough (longer than the sum of the average service times, which is typical in most clinical settings), and there is no variability in service time (i.e., $CV_i = 0$), then the patient length of visit can be calculated by summing up all the service times (since there is no resource availability issue). Consequently, we define such a length of visit as

$$T_i^{fix} = \sum_{j=1}^{4} \frac{1}{c_{j,i}}. \tag{4.17}$$

Then, the length of visit under non-exponential assumptions can be adjusted based on T_i^{fix} by the CVs of service and inter-arrival times. Based on the results of extensive numerical studies, define an effective CV square, $CV^2_{\text{eff},i}$, for room i:

$$CV^2_{\text{eff},i} = \frac{\sum_{j=1}^{4} \frac{CV^2_{j,i}}{c_i}}{\sum_{j=1}^{4} \frac{1}{c_{j,i}}}. \tag{4.18}$$

Then, the length of visit can be adjusted based on T_i^{fix}, T_i^{exp}, and $CV^2_{\text{eff},i}$ using an empirical formula. Specifically, we hypothesize that there exists a linear relationship of lengths of visit between $CV = 0$ and 1 based on numerical investigations. We propose empirical formulas to approximate the patient length of visit in the system, $T_i^{non-exp}$, when inter-arrival time and service times are non-exponential:

$$T_i^{non-exp} = CV^2_{arrival,i}(T_i^{cv} - T_i^{fix}) + T_i^{fix}, \tag{4.19}$$

where

$$T_i^{cv} = CV^2_{\text{eff},i}(T_i^{exp} - T_i^{fix}) + T_i^{fix}, \tag{4.20}$$

and $CV_{arrival,i}$ is the CV of the patient inter-arrival time for exam room i.

Accuracy

Following the proposed method, the patient length of visit in the system with general probability distributions can be estimated. To evaluate the accuracy of the empirical formulas, dozens of simulation experiments have been carried out. Define T_i^{sims} and $T_i^{non-exp}$ as the length of visit obtained by simulations and the empirical formula, respectively. Let ϵ be the relative difference, defined as

$$\epsilon = \frac{T_i^{non-exp} - T_i^{sims}}{T_i^{sims}} \cdot 100\%.$$

From all the randomly generated scenarios, the average of the absolute difference $|\epsilon|$ is less than 3%, with the largest one being within 10%. In Table 4.1 and Figure 4.4, two systems of three-identical exam rooms are highlighted. The mean service and inter-arrival times are fixed throughout the tests. The CVs of the service and arrival distributions are chosen from [0,1] in an ascending sequence from tests 1 through 9 for each system. In both figures, the results from simulations and the empirical formulas are represented by red crosses and blue squares, respectively. As one can see, the differences in lengths of visit between the results obtained from simulations and the empirical formulas are within 5%, and $T^{non-exp}$ is within the 95% confidence interval of T^{sims} in most of the cases. Therefore, we conclude that the empirical formulas (4.17) - (4.20) provide an accurate estimate of the length of visit in most cases when CVs take value within 0 to 1.

4.4 Conclusions

In this chapter, based on the analytical modeling framework introduced in Chapter 3, an iterative method is proposed to address the dimensionality issue for the care delivery system with shared resources. An accurate estimation of the patient length of visit is achieved. The model introduced in this chapter can be applied to study various care delivery systems that share similar characteristics such as multi-stages of services, limited care providers, and multiple patient rooms. The model provides hospital/clinic professionals a quantitative tool to evaluate the current system performance, investigate the effects of different configurations, and predict the operational efficiency for future plans, which is critical for improving the decision making in healthcare operations management.

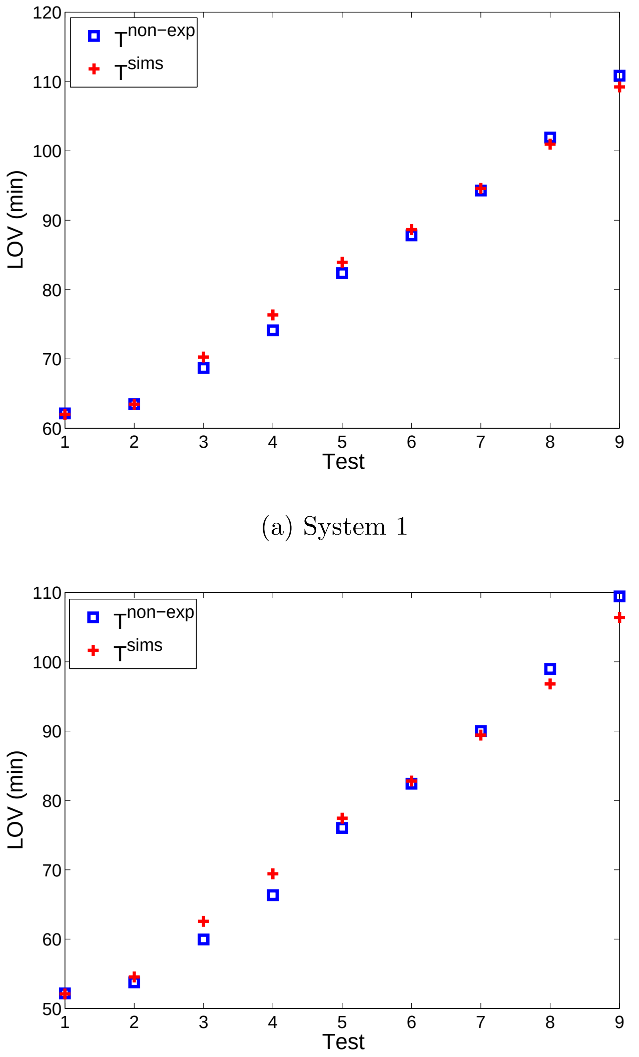
Table 4.1: Accuracy of the empirical formula for estimating LOV

(a) System 1

test	$T^{non-exp}$	T^{sims}	95% CI	ϵ (%)
1	62.14	62.01	(61.91, 62.11)	0.21
2	63.46	63.46	(63.21, 63.72)	-0.00
3	68.69	70.26	(69.54, 70.99)	-2.23
4	74.12	76.34	(75.36, 77.33)	-2.92
5	82.37	83.93	(82.61, 85.25)	-1.87
6	87.81	88.64	(87.19, 90.09)	-0.94
7	94.28	94.58	(92.51, 96.64)	-0.31
8	101.92	100.96	(98.69, 103.23)	0.95
9	110.84	109.22	(106.17, 112.26)	1.48

(b) System 2

test	$T^{non-exp}$	T^{sims}	95% CI	ϵ (%)
1	52.18	52.10	(52.03, 52.17)	0.14
2	53.77	54.55	(54.33, 54.76)	-1.43
3	59.95	62.57	(61.86, 63.26)	-4.17
4	66.34	69.42	(68.32, 70.51)	-4.43
5	76.03	77.45	(76.12, 78.77)	-1.82
6	82.42	82.79	(81.35, 84.22)	-0.44
7	90.02	89.43	(87.27, 91.58)	0.65
8	98.97	96.80	(94.25, 99.34)	2.23
9	109.42	106.37	(101.47,111.27)	2.86



(a) System 1

(b) System 2

Figure 4.4: LOV comparison between simulation and empirical estimation

Chapter 5

Modeling of Primary Care Delivery with E-Visits

5.1 Introduction

Many healthcare organizations have initiated e-visit programs to provide patient-physician communications through securing messages. In this chapter, we introduce a quantitative model to study primary care delivery with e-visits. In Section 5.2, system descriptions and the structural modeling are introduced. In Section 5.3, analytical formulas to evaluate the average patient length of visit and its variance in primary care clinics with e-visits are derived. System-theoretic properties are investigated and different operating policies coordinating office and e-visits are compared in Section 5.4.

5.2 System Description and Modeling

As depicted in Figure 5.1, care delivery process is essentially a service network and patients can get access to care through different venues: web service, which is usually for patients to inquire some standard questions about simple diseases through an online questionnaire program; e-visits, mainly for the patients with low-acuity complaints and ongoing care of chronic diseases to communicate with physicians; office visits, traditional face-to-face encounters; urgent care, for after hour visits or walk-in for a quick treatment, where scheduling is not required; emergency department, for night and emergent visits. After completing the online programs, a patient may still seek communication with his/her primary care physician through e-visit if the online evaluation is not sufficient or satisfactory. In addition, the support staff will review the web service results and, if needed, forward those complex inquiries to patient's primary care physician for a follow-up e-visit. Therefore, patients can transfer from web service to e-visit. Similarly, after e-visit, a patient might still need an office visit according to his/her health status and the complication of the disease. In the case of long queues or extended waiting time for office visits, or during after hours, patients may seek care services through other channels such as urgent care units, and if not available, emergency departments for prompt treatment.

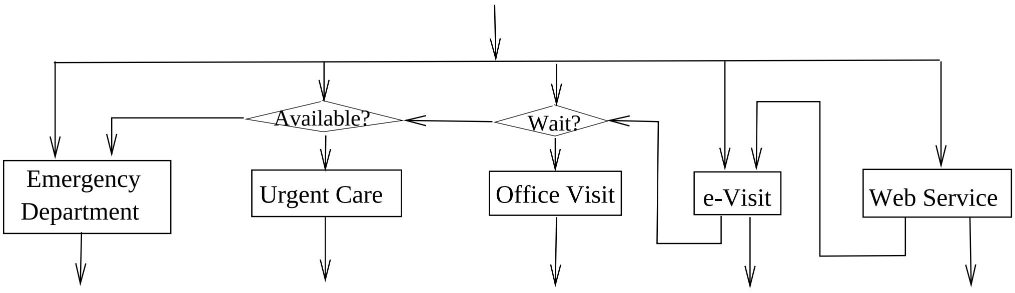


Figure 5.1: Patient flow in care delivery systems

As the focus of this chapter is on studying e-visit and its impact on primary care physicians' operations, only web service, e-visit and office visit are considered (see Figure 5.2). The majority of patients in primary care clinics are associated with their primary care physicians. Therefore, due to physician-patient match, we consider a model with all the services linked with one physician. Assumptions below address the patients, the services, and their interactions.

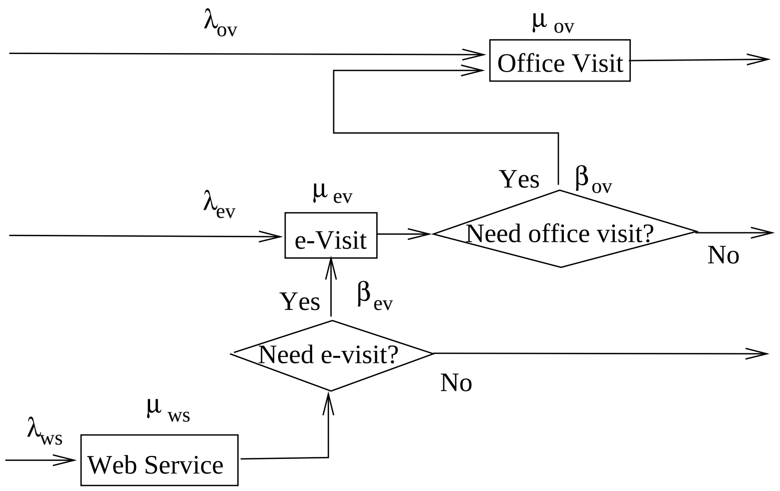


Figure 5.2: Network model for primary care patient flow

1) The patients associated with the same primary care physician access care services with the following Poisson arrival rates: λ_{ws} for web services, λ_{ev} for e-visits, and λ_{ov} for office visits.

2) The primary care physician's service times for e-visits and office visits are described by probability distributions with service rates μ_{ev} and μ_{ov}, coefficients of variation (CVs) cv_{ev} and cv_{ov}, as well as the the third moments (skewness) $E(S_{ev}^3)$ and $E(S_{ov}^3)$, correspondingly.

3) After web service, a patient has the probability β_{ev} to seek an e-visit for further inquiries. After e-visit, a patient may need to go for an office visit with the probability β_{ov}.

4) The physician also deals with billings and documentations intermediately between patient visits. When no patient is waiting, he/she works on non-direct care related tasks, and the duration of tasks follows a probability distribution with the vacation

rate μ_v, the CV cv_v, and third moment $E(S_v^3)$. The physician will return to serve patients only after finishing an ongoing activity.

5) The following scheduling policies for coordinating office and e-visits are proposed: (a) non-preemptive, i.e., an ongoing e-visit service will not be interrupted even if an office visit patient arrives; (b) preemptive-resume, i.e., the current e-visit service can be interrupted if an office visit patient arrives, and the e-visit will resume afterward (in both policies, office visit has a higher priority); (c) first come first serve, i.e., the service will be carried out without priority but only based on who comes earlier.

In an appropriately defined state space, the system described by i)-v) forms a stationary random process. To quantify the system performance, currently we only focus on time-related measurements and do not consider other factors. Intuitively, an efficient system will lead to a shorter cycle time and less waiting for patients and, therefore, it's natural to use cycle time (or patient length of visit) to evaluate the operational efficiency of the system. However, a desired mean time performance alone cannot guarantee patient satisfaction. If the system variation is large, patients may wait for an extremely long time even the mean waiting time is moderate, and the unexpected variation may also impact the clinical outcome and patient safety. Therefore, evaluating the variability of the patient length of visit is also important. Let T_i and Var_i denote the mean and variance of patient length of visit for the type i service, and $i = ev, ov$, representing e-visits and office visits. In the framework of i)-v), T_i and Var_i are functions of all system parameters

$$T_i = f_{T,i}(\mathcal{L}, \mathcal{M}, \mathcal{B}, \mathcal{CV}), \quad i = ev, ov, \tag{5.1}$$

$$\text{Var}_i = f_{\text{Var},i}(\mathcal{L}, \mathcal{M}, \mathcal{B}, \mathcal{CV}, \mathcal{E}), \quad i = ev, ov, \tag{5.2}$$

where

$$\begin{aligned}
\mathcal{L} &= [\lambda_{ws}, \lambda_{ev}, \lambda_{ov}], \\
\mathcal{M} &= [\mu_{ev}, \mu_{ov}, \mu_{v}], \\
\mathcal{B} &= [\beta_{ev}, \beta_{ov}], \\
\mathcal{CV} &= [cv_{ev}, cv_{ov}, cv_{v}], \\
\mathcal{E} &= [E(S_{ev}^3), E(S_{ov}^3), E(S_{v}^3)].
\end{aligned}$$

Remark 5.1 In addition to serving office and e-visit patients, physicians work on other tasks which are not directly encountering patients, such as documentation, paperwork, and dealing with insurance and billings. Assumption iv) implies that the physician works on these activities whenever no patients are waiting. When a new patient arrive, he/she will go back to serve that patient after finishing the current activity.

The problem addressed in this chapter is: *Under assumptions i)-v), develop a method to evaluate the mean and variance of patient length of visit, and investigate system properties and the impact of different scheduling policies between the office and the e-visits.*

The solutions to this problem are presented in Sections 5.3- 5.4.

5.3 Performance Evaluation

5.3.1 Average Length of Visit

Consider the primary care physician's operations described in Section 5.2. For e-visit patients, the arrival includes patients directly seeking e-visits and those coming to e-visits after web services, which is characterized by the transition probability β_{ev}. Thus, the effective arrival rate for e-visits is

$$\lambda'_{ev} = \lambda_{ev} + \beta_{ev}\lambda_{ws}. \tag{5.3}$$

Similarly, the actual arrival rate for office visits also includes the patients who directly request office visits and those seeking face-to-face encounters after e-visits, which has the probability β_{ov}. Thus

$$\lambda'_{ov} = \lambda_{ov} + \beta_{ov}\lambda'_{ev}. \tag{5.4}$$

Remark 5.2 Note that, here, we assume the web service patients who require additional care will first seek e-visits then, if needed, office visits. Clearly, there exists the possibility that they may directly go for office visits. Then, (5.4) can be adjusted as follows:

$$\lambda'_{ov} = \lambda_{ov} + \beta_{ov}\lambda'_{ev} + \beta'_{ov}\lambda_{ws},$$

where β'_{ov} is the referral ratio from the web service to office visits.

Since we focus on one physician and his/her patients, we can model the physician's activities as a single server working on two types of patients: office and e-visits. The non-direct care activities carried out by the physician when no patients are waiting can be modeled as "vacations" with random vacation time. In addition, as the primary care physician offers both office and e-visit services, it is quite common that he/she may expect both types of patients waiting to be served. Then how he/she schedules the work is of interest. Here, we consider three scheduling policies (assumption (v)): non-preemptive, preemptive-resume, and first come first serve.

To derive the patient length of visit under non-preemptive and preemptive-resume policies, we first consider the case without vacations to obtain the patient waiting time as a function of the residual time serving each type of patients. Then, to incorporate vacations, the residual time is modified by accounting for vacations when the server is idle. In this way, the adjusted waiting time is obtained and the patient length of visit can be calculated. In the case of the first come first serve policy, a new probability distribution can be constructed to model the physician's services. Similar approaches are applied to calculate the residual time and the waiting time.

To simplify the derivation and expressions, we introduce the following notations:

$$\rho_i = \frac{\lambda'_i}{\mu_i}, \qquad i = ev, ov, \tag{5.5}$$

$$\rho = \rho_{ev} + \rho_{ov}, \tag{5.6}$$

$$\delta_i = \frac{1 + cv_i^2}{2}, \qquad i = ev, ov, v, \tag{5.7}$$

$$\tau_i = \frac{1}{\mu_i}, \qquad i = ev, ov, v, \tag{5.8}$$

$$\omega_i = \tau_i \delta_i = \frac{E(S_i^2)}{2E(S_i)}, \qquad i = ev, ov, v, \tag{5.9}$$

where ρ and ρ_i's characterize the server utilization; and for service type i, δ_i indicates the variability, τ_i represents the average time, and ω_i is a function of both the mean and the variability. In fact, ω_i represents the ratio between the second and the first moments, multiplied by a factor of 0.5. In the case of the exponential distribution, $\omega_i = \tau_i$, this variable can be explicitly explained by the mean service/vacation time.

Adopting these notations, as well as λ'_{ev} and λ'_{ov} introduced in (5.3) and (5.4), Theorem 5.1 provides the formulas to evaluate the average length of visit for office and e-visit patients with random service and vacation times.

Theorem 5.1 *Under assumptions i)-v), patient average lengths of visit for office and* *e-visit encounters can be calculated as follows:*

$$T_{ev} = \begin{cases} \frac{\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev}+(1-\rho)\omega_v}{(1-\rho_{ov})(1-\rho)} + \tau_{ev}, \\ \qquad \textit{non-preemptive policy}, \\ \frac{\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev}+(1-\rho)\omega_v}{(1-\rho_{ov})(1-\rho)} + \frac{\tau_{ev}}{1-\rho_{ov}}, \\ \qquad \textit{preemptive-resume policy}, \\ \frac{\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev}}{1-\rho} + \omega_v + \tau_{ev}, \\ \qquad \textit{first come first serve policy}, \end{cases} \tag{5.10}$$

$$T_{ov} = \begin{cases} \frac{\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev}+(1-\rho)\omega_v}{1-\rho_{ov}} + \tau_{ov}, \\ \qquad \textit{non-preemptive policy}, \\ \frac{\rho_{ov}\omega_{ov}+(1-\rho)\omega_v}{1-\rho_{ov}} + \tau_{ov}, \\ \qquad \textit{preemptive-resume policy}, \\ \frac{\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev}}{1-\rho} + \omega_v + \tau_{ov}, \\ \qquad \textit{first come first serve policy.} \end{cases} \tag{5.11}$$

Proof: See the Appendix. ■

When the service time and vacation time are exponentially distributed, all the values of cv_i's are equal to 1. The expressions for office and e-visit patients' average lengths of visit can be simplified.

Corollary 5.1 *Under assumptions 1)-5) with exponential service and vacation time distributions, patients' average lengths of visit for office and e-visit encounters can be calculated as*

$$T_{ev}^{exp} = \begin{cases} \frac{\rho_{ov}\tau_{ov}+[1-\rho_{ov}(2-\rho)]\tau_{ev}+(1-\rho)\tau_v}{(1-\rho_{ov})(1-\rho)}, \\ \qquad \textit{non-preemptive policy}, \\ \frac{\rho_{ov}\tau_{ov}+(1-\rho_{ov})\tau_{ev}+(1-\rho)\tau_v}{(1-\rho_{ov})(1-\rho)}, \\ \qquad \textit{preemptive-resume policy}, \\ \frac{\rho_{ov}\tau_{ov}+(1-\rho_{ov})\tau_{ev}}{1-\rho} + \tau_v, \\ \qquad \textit{first come first serve policy}, \end{cases} \tag{5.12}$$

$$T_{ov}^{exp} = \begin{cases} \frac{\tau_{ov}+\rho_{ev}\tau_{ev}+(1-\rho)\tau_v}{1-\rho_{ov}}, \\ \qquad \textit{non-preemptive policy}, \\ \frac{\tau_{ov}+(1-\rho)\tau_v}{1-\rho_{ov}}, \\ \qquad \textit{preemptive-resume policy}, \\ \frac{(1-\rho_{ev})\tau_{ov}+\rho_{ev}\tau_{ev}}{1-\rho} + \tau_v. \\ \qquad \textit{first come first serve policy}. \end{cases} \tag{5.13}$$

Proof: By plugging in $\omega_i = \frac{1}{\mu_i} = \tau_i$, $i = ev, ov, v$, the expressions in (5.12) and (5.13) are obtained after several steps of algebraic operations. ■

Theorem 5.1 and Corollary 5.1 provide formulas to evaluate the length of visit for office and e-visit patients under different scheduling policies.

5.3.2 Variance of Length of Visit

In this subsection, the variances of office and e-visit patients' lengths of visit under the three scheduling policies introduced in assumption 5) are evaluated.

Theorem 5.2 *Under assumptions 1)-5), the variance of patient length of visit can be calculated as follows:*

- *non-preemptive policy*

$$\begin{aligned} Var_{ov} &= \frac{2\delta_{ov}-1}{\mu_{ov}^2} + \frac{\rho_{ov}^2\omega_{ov}^2 - [(1-\rho)\omega_v + \rho_{ev}\omega_{ev}]^2}{(1-\rho_{ov})^2} \\ &+ \frac{(1-\rho)E(S_v^3)\mu_v + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho_{ov})}, \end{aligned} \tag{5.14}$$

$$\begin{aligned} Var_{ev} &= \frac{2\delta_{ev}-1}{\mu_{ev}^2} + \frac{(1-\rho)\omega_v + \rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev}}{(1-\rho_{ov})^2(1-\rho)} \cdot \Big(\frac{2\rho_{ov}\omega_{ov}}{1-\rho_{ov}} + \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev}}{1-\rho} - \omega_v\Big) \\ &+ \frac{(1-\rho)E(S_v^3)\mu_v + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho_{ov})^2(1-\rho)}; \end{aligned} \tag{5.15}$$

- *preemptive-resume policy*

$$\begin{aligned} Var_{ov} &= \frac{2\delta_{ov}-1}{\mu_{ov}^2} + \frac{\rho_{ov}^2\omega_{ov}^2 - (1-\rho)^2\omega_v^2}{(1-\rho_{ov})^2} \\ &+ \frac{(1-\rho)E(S_v^3)\mu_v + \lambda'_{ov}E(S_{ov}^3)}{3(1-\rho_{ov})}, \end{aligned} \tag{5.16}$$

$$\begin{aligned} Var_{ev} &= \frac{2\delta_{ev}-1}{\mu_{ev}^2(1-\rho_{ov})^2} + \frac{(1-\rho)\omega_v + \rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev}}{(1-\rho_{ov})^2(1-\rho)} \cdot \Big(\frac{2\rho_{ov}\omega_{ov}}{1-\rho_{ov}} + \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev}}{1-\rho} - \omega_v\Big) \\ &+ \frac{(1-\rho)E(S_v^3)\mu_v + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho_{ov})^2(1-\rho)}; \end{aligned} \tag{5.17}$$

- *first come first serve policy*

$$\begin{aligned} Var_j &= \frac{2\delta_j-1}{\mu_j^2} + \frac{(\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev})^2}{(1-\rho)^2} - \omega_v^2 + \frac{\mu_v}{3}E(S_v^3) \\ &+ \frac{\lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho)}, \qquad j = ev, ov. \end{aligned} \tag{5.18}$$

Proof: See the Appendix. ■

As one can see, in addition to the mean and CV of service and vacation times, the third moments play a role in determining system performance variations. When exponential service and vacation time distributions are assumed, all CVs are equal to 1 and the expressions for the third moments can be explicitly expressed by parameter μ_j's.

Corollary 5.2 *Under assumptions 1)-5), the variance of patient length of visit under exponential service and vacation time distributions can be calculated as:*

- *non-preemptive policy*

$$\begin{aligned} Var_{ov}^{exp} &= \frac{1}{\mu_{ov}^2} - \frac{\rho_{ev}^2}{(1-\rho_{ov})^2\mu_{ev}^2} + \frac{\rho_{ov}^2}{(1-\rho_{ov})^2\mu_{ov}^2} + \frac{(1-\rho)(1+\rho_{ev}-\rho_{ov})}{(1-\rho_{ov})^2\mu_v^2} \\ &+ \frac{2\rho_{ov}}{(1-\rho_{ov})\mu_{ov}^2} + \frac{2\rho_{ev}}{(1-\rho_{ov})\mu_{ev}^2} - \frac{2\rho_{ev}(1-\rho)}{(1-\rho_{ov})^2\mu_{ev}\mu_v}, \qquad (5.19) \\ Var_{ev}^{exp} &= \frac{1}{\mu_v^2(1-\rho_{ov})^2} + \frac{\rho_{ev}^2}{\mu_{ev}^2(1-\rho)^2(1-\rho_{ov})^2} + \frac{2\rho_{ev}}{\mu_{ev}^2(1-\rho)(1-\rho_{ov})^2} \\ &+ \frac{\rho_{ov}^2(3-2\rho-\rho_{ov})}{\mu_{ov}^2(1-\rho)^2(1-\rho_{ov})^3} + \frac{2\rho_{ov}}{\mu_v\mu_{ov}(1-\rho_{ov})^3} + \frac{2\rho_{ov}}{\mu_{ov}^2(1-\rho)(1-\rho_{ov})^2} \\ &+ \frac{2\rho_{ov}\rho_{ev}(2-\rho_{ov}-\rho)}{\mu_{ov}\mu_{ev}(1-\rho)^2(1-\rho_{ov})^3} + \frac{1}{\mu_{ev}^2}; \qquad (5.20) \end{aligned}$$

- *preemptive-resume policy*

$$\begin{aligned} Var_{ov}^{exp} &= \frac{\rho_{ov}^2}{\mu_{ov}^2(1-\rho_{ov})^2} + \frac{(1-\rho)(1+\rho_{ev}-\rho_{ov})}{\mu_v^2(1-\rho_{ov})^2} + \frac{2\rho_{ov}}{\mu_{ov}^2(1-\rho_{ov})} \\ &+ \frac{1}{\mu_{ov}^2}, \qquad (5.21) \\ Var_{ev}^{exp} &= \frac{1}{\mu_v^2(1-\rho_{ov})^2} + \frac{\rho_{ev}^2}{\mu_{ev}^2(1-\rho)^2(1-\rho_{ov})^2} + \frac{2\rho_{ev}}{\mu_{ev}^2(1-\rho)(1-\rho_{ov})^2} \\ &+ \frac{\rho_{ov}^2(3-2\rho-\rho_{ov})}{\mu_{ov}^2(1-\rho)^2(1-\rho_{ov})^3} + \frac{2\rho_{ov}}{\mu_v\mu_{ov}(1-\rho_{ov})^3} + \frac{2\rho_{ov}}{\mu_{ov}^2(1-\rho)(1-\rho_{ov})^2} \\ &+ \frac{2\rho_{ov}\rho_{ev}(2-\rho_{ov}-\rho)}{\mu_{ov}\mu_{ev}(1-\rho)^2(1-\rho_{ov})^3} + \frac{1}{(1-\rho_{ov})^2\mu_{ev}^2}; \qquad (5.22) \end{aligned}$$

- *first come first serve policy*

$$Var_j^{exp} = \frac{1}{\mu_j^2} + \frac{1}{\mu_v^2} + \frac{2(\frac{\rho_{ov}}{\mu_{ov}^2} + \frac{\rho_{ev}}{\mu_{ev}^2})}{(1-\rho)} + \frac{(\frac{\rho_{ov}}{\mu_{ov}} + \frac{\rho_{ev}}{\mu_{ev}})^2}{(1-\rho)^2},$$

$$j = ov, ev. \tag{5.23}$$

Proof: By plugging in $\delta_i = 1$, $\omega_i = \frac{1}{\mu_i}$, $E(S_i^3) = \frac{6}{\mu_i^3}$, $i = ev, ov, v$, the above expressions can be obtained after several steps of algebraic operations. ■

Building upon these system performance evaluation formulas, system properties like monotonicity can be studied. Then, questions such as how do system parameters impact performance measures and what are the directions to improve system performance can be answered. In the Sections 5.4-5.6, the properties of the mean and variance of length of visit are discussed, and different scheduling policies are compared.

5.4 System Properties

5.4.1 Property of Average Length of Visit

In this section, we investigate the impact of routing probabilities on e-visit and office visit patients' average lengths of visit. Since β_{ev} and β_{ov} are the probabilities that patients continue to seek e-visits and office visits after web service and e-visits, respectively, the monotonicity of T_i, $i = ev, ov$, with respect to them could provide insights on how e-visits impact patient access to primary care.

Monotonicity of T_{ov} with respect to β_{ov}

Proposition 5.1 *Under assumptions 1)-5), T_{ov} is monotonically increasing with respect to β_{ov}, i.e., $\frac{\partial T_{ov}}{\partial \beta_{ov}} > 0$, if and only if*

$$\begin{cases} \omega_{ov} > (\omega_v - \omega_{ev})\rho_{ev}, & \textit{non-preemptive policy,} \\ \omega_{ov} > \omega_v \rho_{ev}, & \textit{preemptive-resume policy,} \\ \textit{without condition,} & \textit{first come first serve policy.} \end{cases}$$

Proof: See the Appendix. ■

Intuitively, if the routing probability of seeking office visits after e-visits, β_{ov}, is increasing, the physician's workload with office visit patients is increasing. Under the non-preemptive policy, when $\omega_{ov} > (\omega_v - \omega_{ev})\rho_{ev}$, the office visit patient's length of visit will increase with respect to β_{ov}, and will be non-increasing vice versa. Such a condition suggests that, roughly, the moment ratio of the office service is larger than that of the difference between vacation and e-visit.

In practice, this type of condition typically holds, since both e-visits and vacations have lower priority than office visits and usually take a shorter time compared with office visits. The difference will be even smaller considering the discount factor $\rho_{ev} < 1$. In particular, when service and vacation times are exponentially distributed, this condition is simplified to $\tau_{ov} > (\tau_v - \tau_{ev})\rho_{ev}$, which again holds most of the time.

Under the preemptive-resume policy, the condition becomes more strict, where $\omega_{ov} > \omega_v \rho_{ev}$ (in the exponential case, $\tau_{ov} > \tau_v \rho_{ev}$) is required. The reason is that under the preemptive-resume policy, the physician will stop working on e-visit patient and immediately serve an incoming office visit patient. Then, the service time and the variability of e-visits will not play a significant role in the waiting time of office visit patients compared with the non-preemptive case, where the physician has to finish any ongoing e-visit service before moving to office visit patients. However, since vacations usually take a shorter time and $\rho_{ev} < 1$, this condition is typically satisfied, so that the monotone increasing property holds.

An illustration of such monotonicity property in exponential scenarios is exhibited in Figure 5.3, in which the parameters are selected as follows:

$$\beta_{ov} \in [0, 0.5), \qquad \beta_{ev} = 0.5, \tag{5.24}$$

$$\text{Case A:} \qquad \tau_v = 30\tau_{ev} = 10\tau_{ov}, \tag{5.25}$$

$$\text{Case B:} \qquad \tau_v = \tau_{ev} = \frac{\tau_{ov}}{3}. \tag{5.26}$$

The reason to include the seldom occurring Case A is to show the decreasing monotonicity. As one can see, when office visits take a longer time, which meets the requirement in Proposition 5.1, T_{ov} is increasing with respect to β_{ov} (Case B). However, if vacation (or non-direct care) takes an extremely longer time than office and e-visits, T_{ov} could decrease with respect to β_{ov} (Case A). In a sense, waiting for short office visits is better than for long vacations.

When the first come first serve policy is applied, the office visit patient's length of visit is monotonically increasing with respect to β_{ov} without any condition. In this case, both office and e-visits are treated with equal priority. Increasing physician's workload (ρ_{ov} and ρ in (5.11) and (5.13)) will lead to a longer patient length of visit.

Therefore, in most of the practical cases, if more patients need to seek additional office visits after e-visits, the accessibility to office visits can be further impaired. Thus, planning e-visits properly to limit this routing probability is of importance, and will be part of future work.

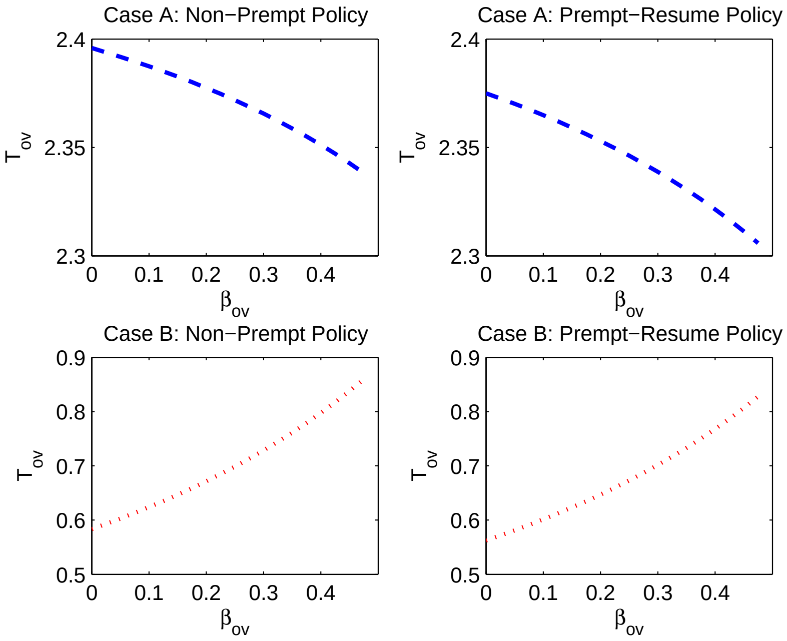


Figure 5.3: Monotonicity of T_{ov} w.r.t. β_{ov}

Monotonicity of T_{ov} with respect to β_{ev}

Proposition 5.2 *Under assumptions 1)-5), T_{ov} is monotonically increasing with respect to β_{ev}, i.e., $\frac{\partial T_{ov}}{\partial \beta_{ev}} > 0$, if and only if*

$$\begin{cases} \beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov} - \lambda_{ov})(\omega_v - \omega_{ev}), \\ \qquad\qquad \textit{non-preemptive policy}, \\ \beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov} - \lambda_{ov})\omega_v, \\ \qquad\qquad \textit{preemptive-resume policy}, \\ \textit{without condition, first come first serve policy.} \end{cases}$$

Proof: See the Appendix. ■

Again, the increasing monotonicity exists without any condition under the first come first serve policy. For non-preemptive and preemptive-resume policies, the necessary and sufficient conditions become more complex.

When β_{ev} is increasing, i.e., more patients continue to seek e-visits after web services, which leads to an increase in the number of patients to further come to the office visit (as $\beta_{ov} > 0$ and is kept constant). Since β_{ev} mainly affects the arrival of e-visits, only when β_{ov} is large enough, the increase of follow-up office visits can exert a significant effect (which explains the conditions with the factor β_{ov} on the left-hand side of the inequalities in Proposition 5.2, required for both the policies).

For the non-preemptive policy, if the physician spends more time, which also has a higher variability on office and e-visits than vacation, a longer length of visit can be observed (which explains the condition regarding the ω_{ov} and $\omega_v - \omega_{ev}$ factors in Proposition 5.2 for the non-preemptive policy). For the preemptive-resume policy, additional e-visit patients will not significantly impact office visits, since the physician will stop working on any e-visit and immediately work on the coming office visit patient. Thus, the condition in Proposition 5.2 for the preemptive-resume policy becomes stricter, where the $\omega_v - \omega_{ev}$ term changes to ω_v.

Note that these conditions are necessary and sufficient, which indicates that if these conditions are not met, T_{ov} will be monotone non-increasing with respect to β_{ev}. Figure 5.4 illustrates such properties in exponential cases. System parameters are selected as in (5.25) and (5.26), but (5.24) is replaced by (5.27) to represent the scenario that web service has a higher referral ratio than e-visits

$$\beta_{ev} \in [0, 0.95), \quad \beta_{ov} = 0.1. \tag{5.27}$$

As exhibited in Figure 5.4, when vacation time is much longer, waiting for more office visits could be even beneficial, so that the decreasing monotonicity can be observed.

Monotonicity of T_{ev}

Unlike T_{ov}, the monotonicity of T_{ev} is consistent for the non-preemptive, preemptive-resume and first come first serve policies.

Proposition 5.3 *Under assumptions 1)-5), T_{ev} is monotonically increasing with respect to β_{ev} and β_{ov}, i.e., $\frac{\partial T_{ev}}{\partial \beta_i} > 0$, $i = ov, ev$.*

Proof: See the Appendix. ■

Proposition 5.3 articulates that the length of visit of e-visit patients is always monotonically increasing with respect to β_{ev} and β_{ov}, no matter which policy is implemented. Larger β_{ov} and β_{ev} increase the effective arrivals, resulting in more patients waiting in line. Under all the policies, a newly arrived e-visit patient needs to wait until all types of patients in line are finished. Thus, the increase of average length of visit can be foreseen.

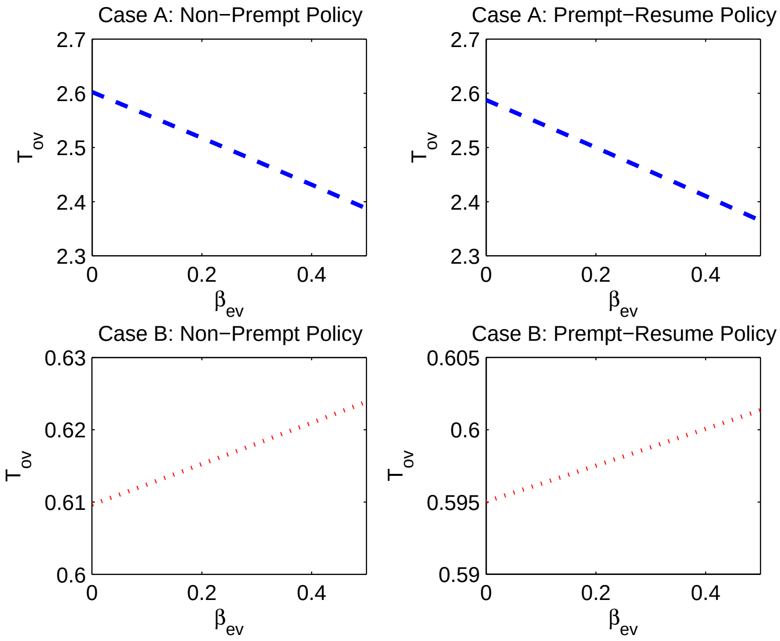


Figure 5.4: Monotonicity of T_{ov} w.r.t. β_{ev}

Discussions

The scheduling policies introduced in this study can be categorized into two groups, based on whether the arriving patients are with or without priority. For the first come first policy, patient types are not differentiated, and increasing either β_{ov} or β_{ev} increases the total patient arrivals, so does the server intensity ρ_{ov}, ρ_{ev}, and ρ. In addition, the effect of vacation on patient length of visit is independent of sever intensity, which is elucidated in (5.10) and (5.11) (where the terms related to ω_v or τ_v are independent of ρ_{ev}, ρ_{ov} and ρ). Therefore, it is straightforward that the increasing monotonicity holds for lengths of visit of both office and e-visit patients unconditionally.

For the policies with priorities, the results differ for office and e-visit patients. As the e-visit patients have a lower priority, their waiting incorporates the waiting for all the patients in line and the waiting for the physician to return from a vacation. Larger β_{ev} or β_{ov} increases the overall patient arrival, and thus the overall number of patients waiting in line. Therefore, the monotonicity of their length of visit holds naturally without conditions.

On the other hand, office visit patients are mainly waiting for office patients in line and the physician returning from a vacation. There exists a tradeoff between waiting for more office and e-visits due to the increase of β_{ov} or β_{ev} and waiting for potentially fewer vacations. Therefore, conditions are required to ensure the monotone increasing of length of visit for office visits. In extreme cases, if vacations are very long or suffer large variations ($\omega_v \gg \omega_{ev}$ or $\omega_v \gg \omega_{ov}$), then having more office and e-visit arrivals could be beneficial (i.e., T_{ov} is monotonically decreasing with respect to β_{ov} and β_{ev}). Moreover, for T_{ov} to be monotonically increasing with β_{ev}, as β_{ev} mainly affects e-visits and its impact on office visit is through β_{ov}, additional conditions on β_{ov} are required.

The conditions for the preemptive-resume policy are always stricter than that for the non-preemptive policy. In the former case, physicians will stop the ongoing e-visit, and thus, only significant changes in e-visits will impose effects on office visits, while in the latter case, physicians will finish the current e-visit service, and any change in e-visits may immediately impact office visits.

In summary, in practical cases, office visits have higher a demand and take a longer time, and then both T_{ov} and T_{ev} are monotonically increasing with respect to β_{ov} and β_{ev}.

5.4.2 Property of Variance of Length of Visit

Monotonicity of Var$_{ov}$

First, we investigate the monotonicity of variance of length of visit Var$_{ov}$ with respect to β_{ov}. The increasing monotonicity holds under a sufficient but not necessary condition.

Proposition 5.4 *Under assumptions 1)-5), Var$_{ov}$ is monotonically increasing with respect to β_{ov}, i.e., $\frac{\partial Var_{ov}}{\partial \beta_{ov}} > 0$, if*

$$\begin{cases} \omega_{ov} \geq \rho_{ev}|\omega_v - \omega_{ev}| \text{ and} \\ \mu_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3) \geq \rho_{ev}\mu_v E(S_v^3), \\ \qquad\qquad\qquad \textit{non-preemptive policy}, \\ \omega_{ov} \geq \rho_{ev}\omega_v \text{ and } \mu_{ov}E(S_{ov}^3) \geq \rho_{ev}\mu_v E(S_v^3), \\ \qquad\qquad\qquad \textit{preemptive-resume policy}, \\ \textit{without condition, first come first serve policy.} \end{cases}$$

Proof: See the Appendix. ■

The sufficient conditions for variance of length of visit are much more complex compared to that of the average length of visit, since the third moments are involved. These conditions indicate that when the office visit has a longer service time and a larger variance, and the vacation (i.e., non-direct care activity) has a smaller moment ratio, then more patients seeking office visits after e-visits will lead to larger variability in the patient flow. Similar to the T_{ov} case, the sufficient conditions under the preemptive-resume policy are stricter than those under the non-preemptive policy. Under the first come first serve policy, fortunately, the monotonicity is straightforward that the variance of length of visit for office visit patients is always increasing when more patients shift to office visits.

It can be noticed that the characteristic of vacation plays an important role in deciding the monotonicity of the mean and variance of patient cycle time. The monotonicity still holds as long as all the three moments of vacation are small enough. One other observation is that the length of visit variation is affected by multiple factors comprising the first, second, and third moments. Therefore, the effect of each factor on determining the holistic system performance weakens. Inference can be drawn that the increase of referral ratio escalates the variation of the system, but not in a strong manner as it impacts the mean time performance. Similar properties are witnessed for the variance with respect to β_{ev}.

These sufficient conditions are less complicated under the exponential assumption.

Corollary 5.3 *Under assumptions 1)-5) with exponential service and vacation time distributions,* $\frac{\partial Var_{ov}^{exp}}{\partial \beta_{ov}} > 0$*, if*

$$\begin{cases} \tau_{ov} \geq \rho_{ev}|\tau_v - \tau_{ev}| \text{ and} \\ \tau_{ov}^2 \geq \rho_{ev}(\tau_v^2 - \tau_{ev}^2), \\ \qquad\qquad \text{non-preemptive policy}, \\ \tau_{ov}^2 \geq \rho_{ev}\tau_v^2, \\ \qquad\qquad \text{preemptive-resume policy}. \end{cases}$$

Proof: By plugging in $\delta_i = 1$, $\omega_i = \frac{1}{\mu_i}$, and $E(S_i^3) = \frac{6}{\mu_i^3}$, $i = ev, ov, v$, and eliminate the redundant conditions, the expressions are obtained. ■

Conditions described above are typically met, since the office visit usually takes a longer time than e-visits and vacations, and $\rho_{ev} < 1$. Illustrations are depicted in Figure 5.5, where the same parameter settings (5.24)-(5.26) matching the conditions in Corollary 5.3 are used. When office visits take a longer time, Var_{ov} increasing with β_{ov} is observed (Case B). However, when the vacation is abnormally long, the decreasing phenomenon can be observed (Case A).

Next we study the monotonicity of Var_{ov} with respect to β_{ev}. Similar to T_{ov}'s case, the sufficient conditions become more strict under non-preemptive and preemptive-resume policies, but no condition is demanded under the first come first policy.

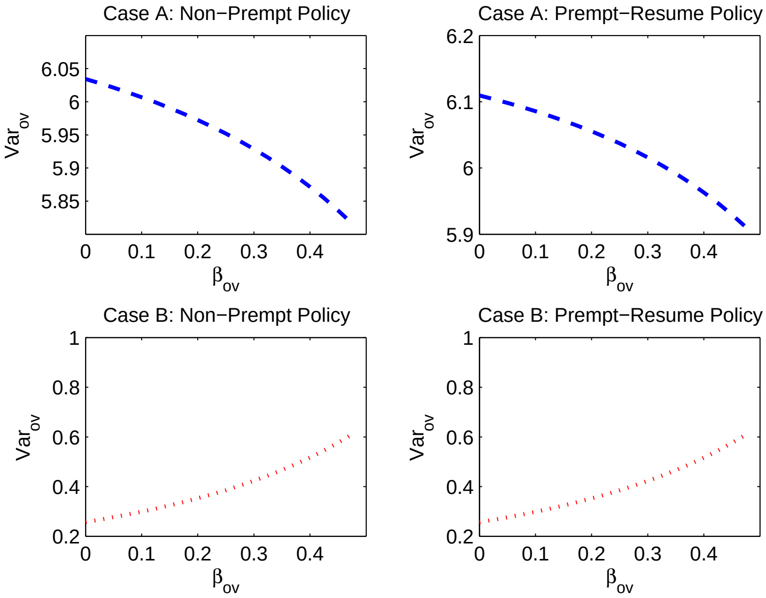


Figure 5.5: Monotonicity of Var$_{ov}$ w.r.t. β_{ov}

Proposition 5.5 *Under assumptions 1)-5), Var_{ov} is monotonically increasing with respect to β_{ev}, i.e., $\frac{\partial Var_{ov}}{\partial \beta_{ev}} > 0$ if*

$$\begin{cases} \beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov}-\lambda_{ov})|\omega_{ev}-\omega_v| \\ \quad \textit{and } \mu_{ev}E(S_{ev}^3) \geq \mu_v E(S_v^3), \\ \qquad\qquad\qquad \textit{non-preemptive policy}, \\ \beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov}-\lambda_{ov})\omega_v \\ \quad \textit{and } \beta_{ov}\mu_{ov}\mu_{ev}E(S_{ov}^3) \geq (\mu_{ov}-\lambda_{ov})\mu_v E(S_v^3), \\ \qquad\qquad\qquad \textit{preemptive-resume policy}, \\ \textit{without condition, first come first serve policy.} \end{cases}$$

Proof: See the Appendix. ■

The sufficient conditions in Proposition 5.5 contain the necessary and sufficient conditions presented in Proposition 5.2. In addition to that, conditions regarding the third moments are required. Likewise, the conditions are more rigid for the preemptive-resume policy compared with the non-preemptive policy. The increasing monotonicity always holds for the first come first serve policy.

In exponential cases, conditions are simplified. Figure 5.6 illustrates the trend that variance changes with referral ratio β_{ev} under exponential settings.

Corollary 5.4 *Under assumptions 1)-5) with exponential service and vacation time distributions,* $\frac{\partial Var_{ov}^{exp}}{\partial \beta_{ev}} > 0$ *if*

$$\begin{cases} \beta_{ov}\mu_{ev}\tau_{ov} > (\mu_{ov}-\lambda_{ov})|\tau_{ev}-\tau_v| \\ \quad \text{and } \tau_{ev}^2 \geq \tau_v^2 \\ \qquad\qquad\qquad \text{non-preemptive policy,} \\ \beta_{ov}\mu_{ev}\tau_{ov} > (\mu_{ov}-\lambda_{ov})\tau_v \\ \quad \text{and } \beta_{ov}\mu_{ev}\tau_{ov}^2 \geq (\mu_{ov}-\lambda_{ov})\tau_v^2, \\ \qquad\qquad\qquad \text{preemptive-resume policy.} \end{cases}$$

Proof: By plugging in $\delta_i = 1$, $\omega_i = \frac{1}{\mu_i}$, and $E(S_i^3) = \frac{6}{\mu_i^3}$, $i = ev, ov, v$, above expressions can be obtained. ■

Monotonicity of Var$_{ev}$

First, we consider non-preemptive and preemptive-resume policies. The monotonicity conditions for Var$_{ev}$ to increase with respect to β_{ov} and β_{ev} become much more complicated. Thus, only the sufficient conditions are pursued. Under both the policies, the

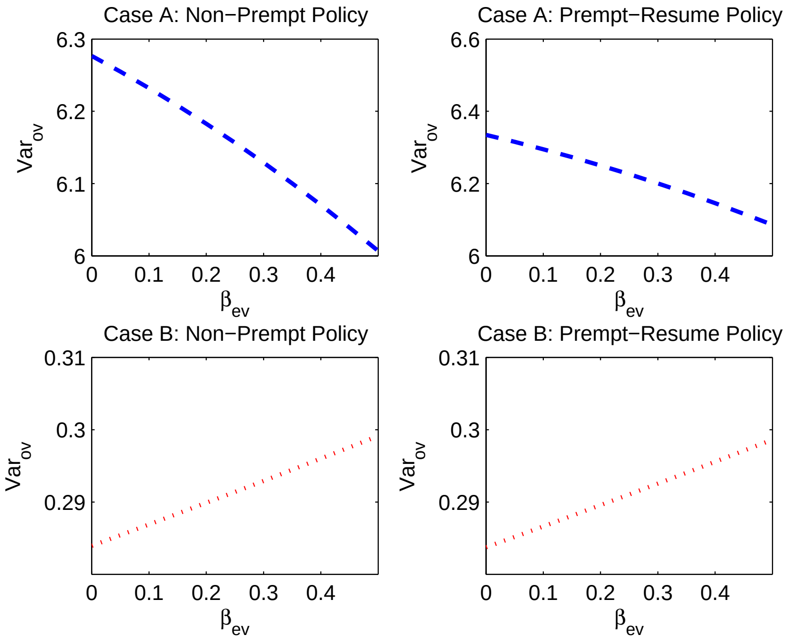


Figure 5.6: Monotonicity of Var_{ov} w.r.t. β_{ev}

same sufficient condition for Var_{ev} to be monotonically increasing with respect to β_{ov} and β_{ev} has been derived.

Proposition 5.6 *Under assumptions 1)-5) with non-preemptive and preemptive-resume policies, if $\omega_{ov} \geq \frac{\omega_v}{2}$ then $\frac{\partial Var_{ev}}{\partial \beta_i} > 0$, $i = ev, ov$, i.e., Var_{ev} is monotonically increasing with respect to β_{ov} and β_{ev}.*

Proof: See the Appendix. ■

Since office visits normally take a longer time and have a larger variability compared with vacations, $\omega_{ov} \geq \frac{\omega_v}{2}$ usually holds and the sufficient condition is satisfied. For the first come first serve policy, monotonic properties with respect to both β_{ov} and β_{ev} hold unconditionally.

Proposition 5.7 *Under the assumptions 1)-5) with the first come first serve policy, Var_{ev} is monotonically increasing with respect to β_{ov} and β_{ev}, i.e.,*

$$\frac{\partial Var_{ev}}{\partial \beta_{ov}} > 0, \qquad \frac{\partial Var_{ev}}{\partial \beta_{ev}} > 0.$$

Proof: See the Appendix. ■

Discussions

For the first come first serve policy, similar to the property of the average length of visit, the terms (the first, second, and third moments) related to vacations are independent of the patient arrival (see (5.18)). Thus, vacations would not affect the system monotonicity property regarding arrival, which is the only term that links to routing probabilities. For the non-preemptive and preemptive-resume policies, the sign of the term ω_v is negative, and the monotonicity of the coefficient of ω_v with respect to β_{ov} or β_{ev} is not clear (see (5.14)-(5.17)). Therefore, the property of vacation plays an important role in determining the monotonicity of variation for office and e-visits.

In summary, the monotone increasing property of variance requires a shorter vacation time, less vacation variations, and a small third moment of vacation. In most practical cases, the variances Var_{ov} and Var_{ev} are monotonically increasing with respect to β_{ov} and β_{ev}, as long as office visits need a longer time and have high variability.

5.5 Comparison of Scheduling Policies

5.5.1 Average Length of Visit

To improve physicians' operations and design an efficient daily workflow, this subsection is dedicated to identifying the optimal scheduling policy and its conditions. By considering all the patients who need physicians' services, the overall patient average length of visit can be obtained, which is a weighted average length of visit of both the office and e-visit patients:

$$\begin{aligned} T &= p_{ev}T_{ev} + p_{ov}T_{ov}, && (5.28)\\ p_i &= \frac{\lambda'_i}{\lambda'_{ev} + \lambda'_{ov}}, \quad i = ev, ov. && (5.29) \end{aligned}$$

Comparing the three scheduling policies, we have Proposition 5.8.

Proposition 5.8 *Under assumptions 1)-5), assume $\mu_{ov} < \mu_{ev}$, then*

$$T_{\textit{Non-Preemp}} > T_{\textit{FCFS}},$$

and in addition, if $cv_{ev} < 1$,

$$T_{\textit{Preemp}} > T_{\textit{Non-Preemp}},$$

where T_{Preemp}, $T_{Non\text{-}Preemp}$, and T_{FCFS} are the weighted average lengths of visit under the preemptive-resume, non-preemptive and first come first serve policies, respectively.

Proof: See the Appendix. ■

The second inequality in Proposition 5.8 indicates that it is always preferable to finish the current e-visit without interruption of the ongoing work, and then start working on office visit patients. The rationale behind this is that the preemptive-resume policy implies that the physician needs to restart the interrupted e-visit after finishing the office visits, which creates more variations for e-visits that can lead to a longer length of visit. In practice, office visits usually take a longer time than e-visits ($\mu_{ov} < \mu_{ev}$), and along with the second inequality, it is justified that the first come first serve policy leads to the highest system productivity, i.e., the shortest average length of visit.

5.5.2 Variance of Length of Visit

The above analysis manifests that the non-preemptive policy is superior to the preemptive-resume policy in terms of mean time performance. To compare the variances of length of visit under different scheduling policies, define the weighted variance of length of visit of both the office and e-visit patients as

$$\text{Var} = p_{ev}\text{Var}_{ev} + p_{ov}\text{Var}_{ov}. \tag{5.30}$$

Then, the following necessary and sufficient condition is obtained.

Proposition 5.9 *Under assumptions 1)-5), if and only if*

$$E(S_{ev}^3) < \frac{1}{(1-\rho_{ov})\mu_{ov}\mu_{ev}^2}(3\rho_{ov} - 6 + \mu_{ev}[12 - 6\rho_{ov} + 6(1-\rho)\mu_{ov}\omega_v]\omega_{ev}$$
$$+ 3\lambda'_{ev}\mu_{ov}\omega_{ev}^2),$$

then

$$Var_{\textit{Non-Preemp}} < Var_{\textit{Preemp}},$$

where $Var_{\textit{Preemp}}$ and $Var_{\textit{Non-Preemp}}$ are the variances of length of visit under the preemptive-resume and non-preemptive policies, respectively.

Proof: See the Appendix. ■

As one can see, if the third moment $E(S_{ev}^3)$ is small enough, then $\mathrm{Var}_{\text{Non-Preemp}}$ is smaller than $\mathrm{Var}_{\text{Preemp}}$. Numerical experiments have demonstrated that such a condition is typically met, so that the non-preemptive policy's superiority holds. In particular, when the exponential service and vacation time distributions are assumed, this conclusion is always valid.

Corollary 5.5 *Under assumptions 1)-5) with exponential service and vacation time distributions, assume $\mu_{ov} < \mu_{ev}$,*

$$Var_{\textit{Non-Preemp}}^{exp} < Var_{\textit{Preemp}}^{exp}.$$

Proof: See the Appendix. ■

The comparison of variances of length of visit between the first come first serve and non-preemptive policies is intricate, even in the exponential case. Although a comparison formula can be derived (see the proof of Proposition 5.9), no simple relationship has been identified. Based on extensive numerical studies, it's observed that the first come first serve policy yields a smaller variance compared with the non-preemptive policy when all the parameters are chosen within the typically range in practice. In particular, under the exponential assumption of service and vacation time distributions, select $\frac{\lambda'_{ov}}{\lambda'_{ev}} \in (1, 11)$, and $\frac{\mu_{ov}}{\mu_{ev}} \in (0.1, 1)$. In addition, let $\rho_{ov}+\rho_{ev} < 1$. A total of 10,000 examples are randomly generated to evaluate the variance. Without a single exception, in all the experiments carried out, it's always the case that

$$\mathrm{Var}^{\mathrm{exp}}_{\mathrm{FCFS}} \leq \mathrm{Var}^{\mathrm{exp}}_{\mathrm{Non\text{-}Preemp}}.$$

Therefore, we recommend the first come first serve policy, which generally leads to a smaller mean and variance of patient length of visit.

5.6 Conclusions

In this chapter, an analytical model of primary care delivery with e-visits has been developed. Formulas to evaluate the mean and variance of office and e-visit patients' average lengths of stay are derived. Three commonly used scheduling policies coordinating office and e-visits are compared and the first come first serve policy is recommended. Such a model builds up a foundation for the further investigation of e-visits' impact on patient access to care.

Chapter 6

Impact of E-Visits on Patient Access to Primary Care

6.1 Introduction

In this chapter, we evaluate the effect of e-visit adoption on patient's accessibility to care. Specifically, we are interested in the conditions under which e-visits should be adopted and, in addition, if e-visits is implemented, how do physicians manage e-visits to improve patient access to primary care. To answer these questions, in Section 6.2, the comparison between systems with and without e-visits is carried out and the criteria for implementing e-visits are identified. In Section 6.3, physician's capacity and the optimal patient diversion to e-visits are investigated. Conclusions are presented in Section 6.4.

For the care delivery system with e-visits, the system performance is evaluated based on the first come first serve policy, which outperforms the other two policies (non-preemptive and preemptive-resume) and generally leads to a smaller mean and variance of patient length of visit. Besides, we mainly focus on the impact of e-visits, and therefore, we do not differentiate direct e-visits and patients seeking e-visits after web services and consequently, $\lambda'_{ev} = \lambda_{ev}$. This change won't affect the conclusions drawn from the previous chapter.

6.2 Comparison between Systems with/without E-Visits

6.2.1 System without E-Visits

Primary care patients typically have different complaints that are of various levels of severity and acuity. Prior to introducing e-visits, no matter with simple care needs or complex complaints, all non-urgent care patients schedule office appointments with their primary care physicians. Here we do not consider patients' urgent care and emergency department visits since they typically are not associated with patients' primary care physicians. Such a unified patient flow is highlighted on the left-hand side of Figure 6.1.

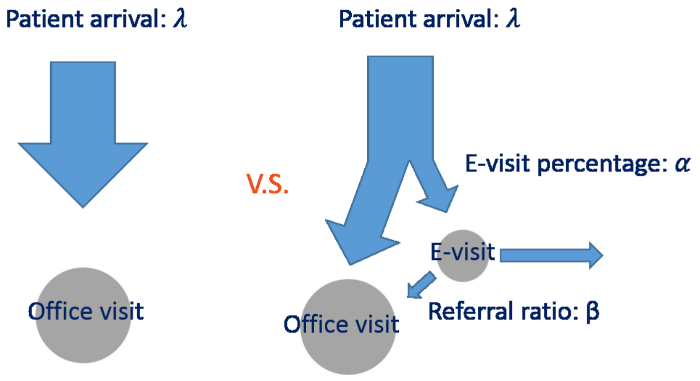


Figure 6.1: Patient flow of systems with and without e-visits

Given a panel which generates patient arrivals with rate λ, the care operations associated with the physician following assumptions described in Chapter 5 can be modeled as a single server queue with server vacations. The physician's utilization (or traffic intensity) on office visits is computed as

$$\rho_t = \frac{\lambda}{\mu_{ov}}. \tag{6.1}$$

Then, we obtain the office visit cycle time T_t under the traditional care delivery model.

$$T_t = \frac{\rho_t \omega_{ov}}{1-\rho_t} + \omega_v + \frac{1}{\mu_{ov}}. \tag{6.2}$$

6.2.2 System with E-Visits

If e-visits is introduced, the new patient flow is illustrated on the right-hand side of Figure 6.1. E-visits function as a diversion for patients with acute non-urgent care needs and patients with chronic disease who need follow-ups. For comparison purposes, we assume the patient panel does not change and the demand remains the same. Since a proportion of patients might be suitable for e-visits, we denote this percentage as α ($0 < \alpha < 1$) and let those patients seek e-visits directly. Then, the arrival rate for e-visits is $\lambda_{ev} = \alpha\lambda$. The rest of patients go for office visits with arrival rate $\lambda_{ov} = (1-\alpha)\lambda$. In addition, some of the e-visit patients might need follow-up office visits. We denote this transition as referral with ratio β_{ov}. Then, the effective arrival rate for office visits λ'_{ov} satisfies

$$\lambda'_{ov} = \lambda_{ov} + \beta_{ov}\lambda_{ev} = (1-\alpha+\alpha\beta_{ov})\lambda. \tag{6.3}$$

To simplify the expressions, we introduce another variable γ, which is the service rate ratio between office and e-visits, $\mu_{ov} = \gamma\mu_{ev}$, and $\gamma < 1$, which implies that e-visits take a shorter service time on average. Besides, the CV of e-visits is no larger than that of office visits, $\delta_{ev} \leq \delta_{ov}$.

Recall that $\rho_t = \frac{\lambda}{\mu_{ov}}$ is the physician utilization when all patients go for office visits (i.e., $\alpha = 0$). When $\alpha \neq 0$, define:

$$\begin{aligned}
\rho_{ov} &= \lambda'_{ov}/\mu_{ov} = (1-\alpha+\alpha\beta_{ov})\rho_t, \\
\rho_{ev} &= \lambda_{ev}/\mu_{ev} = \alpha\gamma\rho_t, \\
\rho &= \rho_{ov} + \rho_{ev} = (1-\alpha(1-\gamma-\beta_{ov}))\rho_t, \qquad (6.4)
\end{aligned}$$

representing physician's utilization on office visits, e-visits, and a total of them, respectively. Considering physician's capacity and to ensure a system that can reach stationary, the physician's utilization satisfies

$$\begin{aligned} \rho_t < 1 \quad &<=> \quad \lambda < \gamma\mu_{ev}, \\ \rho < 1 \quad &<=> \quad (1-\alpha(1-\gamma-\beta_{ov}))\lambda < \gamma\mu_{ev}. \end{aligned} \tag{6.5}$$

Under the same assumptions presented in Section 5.2, for the first come first serve policy, patient average lengths of visit for office and e-visit encounters can be calculated:

$$\begin{aligned} T_{ev} &= (\rho_{ov}\frac{\delta_{ov}}{\mu_{ov}} + \rho_{ev}\frac{\delta_{ev}}{\mu_{ev}})\frac{1}{(1-\rho)} + \omega_v + \frac{1}{\mu_{ev}}, \\ T_{ov} &= (\rho_{ov}\frac{\delta_{ov}}{\mu_{ov}} + \rho_{ev}\frac{\delta_{ev}}{\mu_{ev}})\frac{1}{(1-\rho)} + \omega_v + \frac{1}{\mu_{ov}}. \end{aligned} \tag{6.6}$$

6.2.3 System Comparison

To determine whether the service model with e-visits can outperform the other, we fix the total external arrival rate λ and e-visit diversion factor α. The demand and utilization of provider services are typically shaped by the patient population's age, sex, race, disease burden, etc., which are factors determined by the characteristics of panel patients, but not related to physician's operations.

Many physicians hesitate to adopt e-visits for fear of being overloaded by e-visits as they already bear heavy workload managing office visits. Therefore, the first comparison is regarding physician utilization, which can be captured by "traffic intensity," a measure of how busy the system is.

Proposition 6.1 *Comparing physician's operations before and after the implementation of e-visits, under the condition* $1-\gamma-\beta_{ov} > 0$, $\rho - \rho_t < 0$, *i.e., adopting e-visits could reduce the physician utilization.*

Proof: See the Appendix. ■

Note that $1 - \gamma = \frac{\mu_{ev} - \mu_{ov}}{\mu_{ev}}$, which is the relative service rate difference between office and e-visits. Proposition 6.1 indicates that when patients start using e-visits and the referral ratio is smaller than the relative service rate difference, the physician's workload on serving patients can be reduced. Although being a trivial comparison, the condition identified in Proposition 6.1 is insightful and affects the system performance in other aspects, which will be further discussed in this paper.

Meanwhile, from patients' perspective, for those patients who only receive office visits, we compare the change in cycle time $T_{ov} - T_t$ and conclude that:

Proposition 6.2 *Comparing physician's operations before and after the implementation of e-visits, under the condition* $1 - \gamma - \beta_{ov} \geq 0$*,* $T_{ov} - T_t < 0$*, i.e., adopting e-visits could reduce office visit cycle time.*

Proof: See the Appendix. ■

Furthermore, for those patients who only receive e-visits, we have Corollary 6.1.

Corollary 6.1 *If* $1 - \gamma - \beta_{ov} \geq 0$*, then,* $T_{ev} - T_t < 0$*, i.e., patients experience a shorter cycle time using e-visits compared to e-visit is not offered.*

Proposition 6.2 and Corollary 6.1 suggest that if the efficiency gained from e-visits outweighs the extra workload due to ineffective e-visit usage, although physicians serving office and e-visit patients simultaneously, both office and e-visit cycle times are reduced compared to e-visit is not offered. Note that $T_{ev} < T_{ov}$, since e-visits have a larger service rate. Thus, a sufficient but not necessary condition for $T_{ev} - T_t < 0$ is also $1 - \gamma - \beta_{ov} \geq 0$.

However, it should be noticed that there's a proportion of patients receiving both e-visits and office visits. To further assess whether the whole panel can benefit from e-visit usage, we investigate the overall cycle time for all types of patient visits – a weighted cycle time where the weight is determined by arrival rates. This index is able to incorporate the scenario that some of the e-visit patients still receive office visits afterwards, and therefore, consume more time and resource.

For the system without e-visits, we denote the index as $\phi_{\mathrm{w/o}}$ (without e-visits):

$$\phi_{\mathrm{w/o}} = \lambda T_t = \lambda(\frac{\rho_t \omega_{ov}}{1-\rho_t} + \omega_v + \frac{1}{\mu_{ov}}). \tag{6.7}$$

For the system with e-visits, the index is defined as ϕ_{w} (with e-visits):

$$\phi_{\mathrm{w}} = (1-\beta_{ov})\lambda_{ev}T_{ev} + \beta\lambda_{ev}(T_{ev}+T_{ov}) + \lambda_{ov}T_{ov}. \tag{6.8}$$

Note that $\lambda = (1-\beta_{ov})\lambda_{ev} + \beta_{ov}\lambda_{ev} + \lambda_{ov}$, a redistribution of the total external arrival. Then, apply (6.6) to calculate ϕ_{w}:

$$\phi_{\mathrm{w}} = \rho + (\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} + (1-\rho)\omega_v)\frac{(1+\alpha\beta_{ov})\lambda}{1-\rho}. \tag{6.9}$$

Furthermore, define the system comparison index as $\Phi = \phi_{\mathrm{w/o}} - \phi_{\mathrm{w}}$. If the system with e-visits is more efficient, $\phi_{\mathrm{w}} < \phi_{\mathrm{w/o}}$ and $\Phi > 0$. Otherwise, the system without e-visits is preferred and $\Phi < 0$.

Proposition 6.3 *Comparing physician's operations before and after the implementation of e-visits, if* $1-\gamma-\beta_{ov} \geq 0$, *and the other nondirect care tasks' variation factor* ω_v *satisfies*

$$\omega_v < \frac{\lambda[(-1+2\gamma+(1-\gamma)\gamma\alpha)\delta_{ov} - \gamma^2(1+(1-\gamma)\alpha)\delta_{ev}]}{\gamma(1-\gamma)\mu_{ev}(-\lambda+\gamma\mu_{ev})}, \tag{6.10}$$

then $\Phi > 0$, *i.e., adopting e-visits would decrease the overall cycle time.*

Proof: See the Appendix. ■

Proposition 6.3 provides a sufficient but not necessary condition for $\Phi > 0$. It highlights that for care delivery with e-visits to be more efficient, not only the first moment condition regarding physician's operations is requested, but the variabilities – the relative values among the service time variation factors δ_{ov} and δ_{ev}, and the "vacation" variation factor ω_v all matter. Considering the second moments, we find a criterion opposing e-visit implementation.

Corollary 6.2 *Under the conditions $\delta_{ov} = \delta_{ev}$ and $\beta_{ov} \geq 1 - \gamma$, the index $\Phi < 0$, and adopting e-visits would increase the overall cycle time.*

Proof: See the Appendix. ■

Corollary 6.2 provides a sufficient but not necessary condition to refrain from using e-visits. Basically, when e-visit fails to reduce service variation (we assume $\delta_{ev} \leq \delta_{ov}$), and the efficiency gained cannot compensate the loss of effectiveness, adopting e-visit is not recommended.

As revealed by Proposition 6.3 and Corollary 6.2, there are multiple factors that leverage the value of index Φ, including nondirect care task variation factor ω_v, e-visit service rate μ_{ev}, e-visit service time variation factor δ_{ev}, and e-visit to office visit referral ratio β_{ov}. We further assume these factors are mutually independent and investigate their impact on system efficiency.

Impact of nondirect care tasks

Note that physician's nondirect care work exists no matter e-visit is offered or not. Let the random variable V represent the duration of completing such tasks. $\omega_v = \frac{E(V^2)}{2E(V)}$, which is the half moments ratio of the second moment to the first of "vacation" time. When physician's other tasks are not considered, $\omega_v = 0$. In this case, denote a new comparison index Φ_{nv}.

Proposition 6.4 $\Phi = \Phi_{nv} - \alpha\beta_{ov}\omega_v$. *Compared to* Φ_{nv}*, the performance index of the system incorporating physician's other nondirect care tasks* (Φ) *is smaller.*

Proof: See the Appendix. ■

It can be concluded that physician's other nondirect care tasks do affect the system property. The incorporation of "vacation" essentially offsets the changes e-visits could bring to the overall cycle time. The performance index Φ is decreasing with respect to ω_v. When the other tasks take a very long time or with large variations, the potential benefits from adopting e-visits will be significantly impaired.

On the other hand, even without other tasks, the superiority of having e-visits cannot be guaranteed. In particular, if without "vacation," $\omega_v = 0$, the condition presented in Proposition 6.3 will be mainly determined by the relation between δ_{ov} and δ_{ev}.

Impact of e-visit service time

E-visit service time is characterized by the service rate μ_{ev} and the variation factor δ_{ev}. To investigate the impact of e-visit service time on system comparison, we take derivatives of Φ with respect to μ_{ev} and δ_{ev}, correspondingly.

Proposition 6.5 Φ *is monotonically increasing with respect to e-visit service rate* μ_{ev}*, i.e.,* $\frac{\partial\Phi}{\partial\mu_{ev}} > 0$*. In addition,* Φ *is monotonically decreasing with respect to e-visit service time variation factor* δ_{ev}*, i.e.,* $\frac{\partial\Phi}{\partial\delta_{ev}} < 0$*.*

Proof: See the Appendix. ■

Considering Φ as a function of μ_{ev}, the boundary of Φ with respect to μ_{ev} can be identified. When $\mu_{ev} \to \infty$, i.e., e-visits can be finished instantaneously. Although unlikely, it provides an upper bound of Φ:

$$\begin{aligned}\Phi_{\mu_{ev}\to\infty} &= (1-\beta_{ov})\frac{\alpha\lambda}{\mu_{ov}} - \alpha\beta_{ov}\lambda\omega_v + \frac{\delta_{ov}\lambda^2}{(1-\rho_t)\mu_{ov}^2} \\ &- \frac{\delta_{ov}(1-\alpha+\alpha\beta_{ov})(1+\alpha\beta_{ov})\lambda^2}{(1-\rho_{ov})\mu_{ov}^2}.\end{aligned} \tag{6.11}$$

Next, if $\mu_{ev} \to \mu_{ov}$, then, the lower bound of Φ can be found as

$$\begin{aligned}\Phi_{\mu_{ev}=\mu_{ov}} &= -\frac{\alpha\beta_{ov}\lambda}{\mu_{ov}} - \alpha\beta_{ov}\lambda\omega_v + \frac{\delta_{ov}\lambda^2}{(1-\rho_t)\mu_{ov}^2} \\ &- \frac{(\delta_{ov}(1-\alpha+\alpha\beta_{ov})+\delta_{ev}\alpha)(1+\alpha\beta_{ov})\lambda^2}{(1-\rho)\mu_{ov}^2}.\end{aligned} \tag{6.12}$$

It both cases, the sign of Φ cannot be determined. On one hand, even if physicians can serve e-visit patients very fast, it cannot guarantee the reduction of the overall patient cycle time as long as β_{ov} is large. On the other hand, although e-visits might take approximately the same amount of time as office visits, it still has the potential to reduce cycle time on average if both δ_{ev} and β_{ov} are small. Thus, to improve care delivery, providing standardized and time-saving e-visits is pursued.

Impact of referral ratio

The analysis in Subsection 6.2.3 directs us to investigate how the referral ratio β_{ov} leverages the system performance. Considering Φ as a function of β_{ov} and taking the partial derivative of Φ with respect to β_{ov}, we conclude:

Proposition 6.6 *Φ is monotonically decreasing with respect to the e-visit to office visit referral ratio β_{ov}, i.e., $\frac{\partial\Phi}{\partial\beta_{ov}} < 0$.*

Proof: See the Appendix. ■

Similarly, the range of Φ can be identified based on the monotonicity of Φ with respect to β_{ov}. When $\beta_{ov} = 0$, Φ reaches its maximum:

$$\begin{aligned}\Phi_{\beta_{ov}=0} &= \frac{(1-\gamma)\alpha\lambda}{\gamma\mu_{ev}} + \frac{\alpha\lambda^2((\delta_{ov}-\gamma^2\delta_{ev})\mu_{ev} - (\delta_{ov}-\gamma\delta_{ev})\lambda)}{\mu_{ev}(\lambda-\gamma\mu_{ev})((1-(1-\gamma)\alpha)\lambda-\gamma\mu_{ev})} \\ &> \frac{(1-\gamma)\alpha\lambda}{\gamma\mu_{ev}} + \frac{\alpha\lambda^2(\delta_{ov}-\gamma^2\delta_{ev})(\mu_{ev}-\lambda)}{\mu_{ev}(\lambda-\gamma\mu_{ev})((1-(1-\gamma)\alpha)\lambda-\gamma\mu_{ev})} > 0. \qquad (6.13)\end{aligned}$$

Inequality (6.13) suggests that if e-visit can fully satisfy patients' needs without incurring additional office visits ($\beta_{ov} = 0$), adopting e-visit is beneficial whenever e-visit takes a shorter service time on average ($\gamma < 1$). Such an advantage is even staggering if e-visit consumes significantly less time and with smaller variances according to Proposition 6.5.

Next, when $\beta_{ov} = 1$, the lower bound of Φ is calculated as:

$$\Phi_{\beta_{ov}=1} = -\lambda\{\alpha\omega_v + \frac{\alpha(1-(1-\gamma)\rho_t)\rho_t}{(1-\rho_t)(1-(1+\alpha\gamma)\rho_t)\omega_{ov}} + \frac{\alpha\gamma(1+\alpha)\rho_t\omega_{ev}}{1-(1+\alpha\gamma)\rho_t} + \frac{\alpha}{\mu_{ev}}\} < 0. \quad (6.14)$$

Therefore, when $\beta_{ov} = 1$, the service model with e-visits is not preferred no matter how fast e-visits can be processed. Intuitively, if all e-visit patients seek face-to-face encounters afterwards, then there's no need to provide e-visits which are apparently redundant.

Remark 6.1 Although e-visits provide a gateway for patients with simple complaints, for those patients receiving office visits after e-visits, their office visit time might not be reduced. Therefore, μ_{ov} is kept the same for all office visits. Besides, a referral patient still needs to go through the regular scheduling process for a subsequent office visit.

Numerical experiments

To further study the property of Φ, numerical experiments are carried out. Fix $\lambda = 1$ as the unit arrival rate and ω_v is normalized so the range is set to [0.01, 0.5]. The values of μ_{ev} and γ are randomly generated under the physician utilization constraints (6.5).

The other parameters are chosen from the following sets:

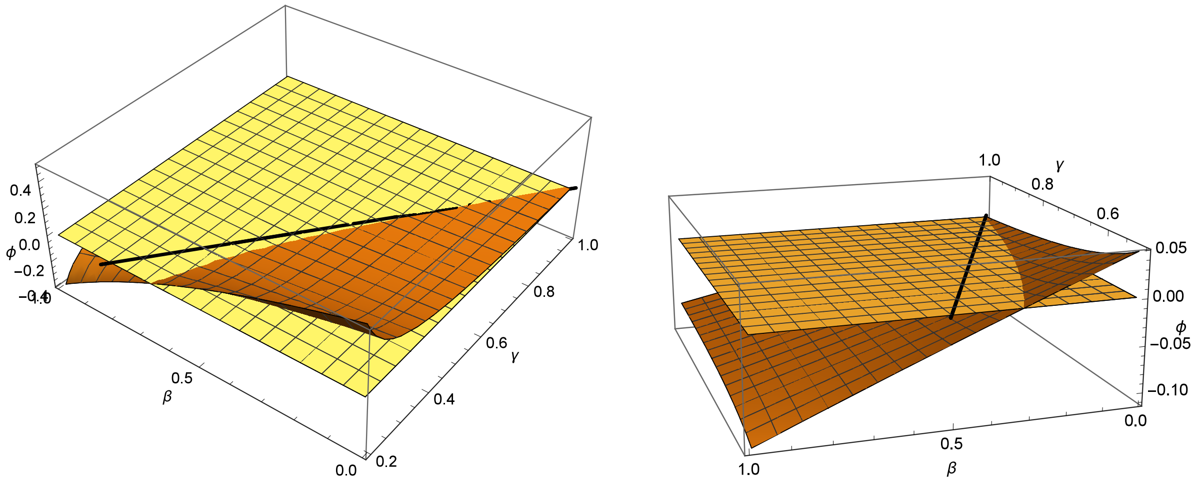
$$\delta_{ev} \in [0.01, 0.99],$$
$$\delta_{ov} \in [0.01, 0.99],$$
$$\alpha \in [0.01, 0.99],$$

and $\delta_{ev} \leq \delta_{ov}$.

Two sample numerical experiments are exhibited in Figures 6.2(a) and 6.2(b). The figures illustrate the three-dimension plot of Φ as a function of β_{ov} and γ. The light color panels represent the plane $\Phi = 0$, and the thick black lines are $\beta_{ov} + \gamma = 1$. They shed light on the trends that when either β_{ov} or γ is large, Φ is small, while when both β_{ov} and γ are small, Φ can be above zero. Moreover, in Figure 6.2(a), the value of ω_v is much smaller than that of Figure 6.2(b). When ω_v is close to zero, the values of β_{ov} and γ that satisfy $\Phi(\beta_{ov}, \gamma) = 0$ are almost coincident with $\beta_{ov} + \gamma = 1$, while when ω_v is large, there's a broad area between the black line and the intersection curve. It corresponds to Proposition 6.3 that physician's longer "vacations" on other tasks would impair the potential benefits from e-visit diversion, so β_{ov} needs to be way smaller than $1 - \gamma$ to make $\Phi > 0$. Practically, other tasks are competing with direct patient care work on physician resource. The "free time" released by e-visits is consumed by nondirect care work, and such influence is significant when nondirect care work is dominant.

6.3 Physician Capacity and Patient Diversion

As we are interested in introducing e-visits to improve patient access, in this section, we draw the attention to two questions. First, what sizes of patient panels are manageable that are compatible with delivering a reasonable level of access to care? Second, how is



(a) Small ω_v

(b) Large ω_v

Figure 6.2: Comparison index Φ as a function of β_{ov} and γ

a manageable patient panel size affected by partial diversion of patient demand to the use of electronic communications?

6.3.1 Physician Capacity Analysis

Viewing the arrival rate (provider visits per day) as a reflection of the physician's panel size, when there're more patients affiliated with the physician, there're more patient complaints that the physician should expect. The previous section investigates the system property by assuming the same total external arrival. In this section, we relax such an assumption. Prior to e-visit, the patient demand a physician could accommodate corresponds to an arrival rate λ_t. When e-visit is implemented, physicians adjust his/her panel size accordingly with an external arrival rate λ. In this subsection, we still fix α and let $0 < \alpha < 1$ since practically, e-visits can only meet limited patient needs.

First, we evaluate the physician's utilization on patient care before (ρ_t) and after (ρ)

e-visit implementation:

$$\rho_t = \frac{\lambda_t}{\gamma\mu_{ev}} < 1, \tag{6.15}$$

$$\rho = \frac{(1-\alpha(1-\gamma-\beta_{ov}))\lambda}{\gamma\mu_{ev}} < 1. \tag{6.16}$$

Proposition 6.7 *Suppose the original arrival rate λ_t results in a reasonable level of physician utilization. Then, when e-visit is implemented, the maximum external arrival rate λ that is manageable (with the physician's utilization unchanged) can be calculated as*

$$\lambda^* = \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}, \tag{6.17}$$

and when $1-\gamma-\beta_{ov} > 0$, $\lambda^* > \lambda_t$, a larger external arrival rate can be accommodated.

Proof: See the Appendix. ■

Conclusions can be drawn that if some of the panel patients use e-visits and the provided e-visit service satisfies the service rate ratio and referral ratio constraint, physicians can potentially accommodate an increased arrival rate and enlarge his/her panel size by approximately $\alpha(1-\gamma-\beta_{ov})100\%$.

Next, it is of interest to us how the system performs when the capacity is expanded. We hereby compare the average cycle times of office and e-visits when $\lambda = \lambda^*$.

Proposition 6.8 *When $\lambda = \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}$ and $1-\gamma-\beta_{ov} \geq 0$, adopting e-visits could reduce both office and e-visit patients' cycle times compared to e-visit is not offered, i.e.,*

$$T_{ov} - T_t < 0, \tag{6.18}$$

$$T_{ev} - T_t < 0. \tag{6.19}$$

Proof: See the Appendix. ■

Furthermore, we take a look at the performance index Φ when $\lambda = \lambda^*$:

$$\begin{aligned}\Phi_{\lambda=\lambda^*} &= \frac{\alpha\lambda_t^2\left(\delta_{ov}(\alpha+\alpha(2\gamma-1)\beta_{ov}+\gamma(\alpha(\gamma-2)+2)-1)-\gamma^2\delta_{ev}(1+\alpha\beta_{ov})\right)}{\gamma\mu_{ev}(1-\alpha(1-\gamma-\beta_{ov}))^2\left(\gamma\mu_{ev}-\lambda_t\right)} \\ &- \frac{\alpha(1-\gamma)\lambda_t\omega_v}{1-\alpha(1-\gamma-\beta_{ov})}. \qquad (6.20)\end{aligned}$$

Unfortunately, the sign of $\Phi_{\lambda=\lambda^*}$ is still difficult to determine even if $1-\gamma-\beta_{ov} \geq 0$ is satisfied. It demonstrates that with the expanded capacity, the reduction in the overall cycle time cannot be guaranteed. Meanwhile, considering the second moments, it can be shown that:

Proposition 6.9 *When $\lambda = \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}$ and $1-\gamma-\beta_{ov} \geq 0$, if $\delta_{ov} = \delta_{ev}$, then, adopting e-visits yields a larger overall cycle time, $\Phi_{\lambda=\lambda^*} < 0$.*

Figure 6.3 demonstrates how the index $\Phi_{\lambda=\lambda^*}$ changes with respect to service time variation factors δ_{ov} and δ_{ev} when the condition $1-\gamma-\beta_{ov} \geq 0$ is satisfied. The thick black line indicates $\delta_{ov} = \delta_{ev}$ and for $\Phi_{\lambda=\lambda^*} > 0$, $\delta_{ov} > \delta_{ev}$ is required.

Therefore, to achieve a larger panel size ($1-\gamma-\beta_{ov} > 0$) and a reduced cycle time ($\Phi_{\lambda=\lambda^*} > 0$) simultaneously, conditions regarding the second moments – the service time and "vacation" time variations are required. In particular, ω_v needs to be small, and $\delta_{ov} > \delta_{ev}$ serves as a necessary but not sufficient condition for $\Phi_{\lambda=\lambda^*} > 0$.

Remark 6.2 Take the partial derivative of Φ with respect to λ:

$$\begin{aligned}\frac{\partial\Phi}{\partial\lambda} &= -\frac{\rho}{\lambda} - (\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev})(\frac{2(1+\alpha\beta_{ov})}{1-\rho} + \frac{(1+\alpha\beta_{ov})\rho}{(1-\rho)^2}) \qquad (6.21) \\ &- (1+\alpha\beta_{ov})\omega_v < 0.\end{aligned}$$

Thus, Φ is monotone decreasing with λ. Ideally, solving for the λ_0 that satisfies $\Phi = 0$ when $\lambda = \lambda_0$, we obtain the threshold of λ for an unchanged overall cycle time performance. However, challenges are that many variables are involved in the expression of λ_0 and a straightforward explanation of λ_0 is hard to achieve.

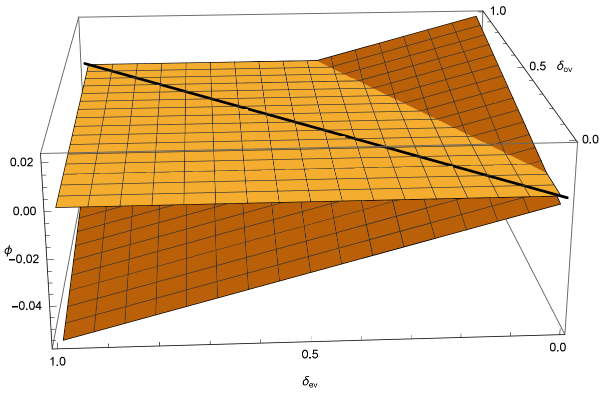


Figure 6.3: Comparison index Φ as a function of δ_{ov} and δ_{ev}

6.3.2 Optimal Patient Diversion

In this subsection, we no longer fix α and explore how a manageable patient panel is affected by the partial diversion of patient demand to e-visits. Proposition 6.7 suggests that λ^* is an increasing function of α. It implies that if the e-visit volume is independent of other variables, under the assumption that $1-\gamma-\beta_{ov} > 0$, it is ideal to have as many patients using e-visits as possible. Nevertheless, this might not be the case in real practice. Based on physicians' experience, among those patient complaints they received, only part of the encounters are suitable for virtual visits without additional referrals. To estimate the proportion of substitutable office visits and determine the optimal diversion α^*, physicians need to understand the characteristics of their patient population. It's important to review community prevalence statistics to assess the symptom categorization, disease burden, and vulnerability and utilization-intensity of their panels. Besides, physicians should also consider patients' individual financial needs, care preferences, and accessibility to e-visit related infrastructure.

Meanwhile, there exists a discrepancy between the ideal diversion as physicians expected and the real usage and acceptance by patients. Some patients advocate e-visits while others might have concerns regarding the quality of care and reimbursement issue, or have limited access to patient portal and internet. Therefore, a match between the "fit" recognized by the physician and patients' reactions is desired. To investigate the proper usage of e-visits, we consider a practical relation between the referral ratio and e-visit volume. Here we follow assumptions (2) and (4) and further assume that the referral ratio increases with respect to e-visit patient volume. The rationale is that if all patients are encouraged to receive e-visits, patients who are not suitable for e-visits might be "mis-triaged" and still need face-to-face encounters, which results in a very high referral ratio and a much longer time to be diagnosed and treated. Specifically, define $\beta_{ov} = k_1\alpha + k_2$, where $k_1, k_2 > 0$ and $k_1 + k_2 < 1$. Then, β_{ov} is within 0 and 1 and is positively correlated with α. On one hand, k_2 represents the fixed effect that regardless of patients making appropriate decisions to seek e-visits or not, some patients still need further assessment that cannot be accomplished via message exchange. On the other hand, k_1 represents the effect that associates with e-visit volume. Having patients with complex complaints treated by e-visits would potentially increase the risk of making office visit referrals. The maximum arrival rate λ^* in Proposition 6.7 indicates that both α and β_{ov} affect the expected new capacity. Substituting β_{ov} by $k_1\alpha + k_2$,

$$\frac{\lambda^*}{\lambda_t} = \frac{1}{1 - \alpha(1 - \gamma - k_1\alpha - k_2)}. \tag{6.22}$$

Figure 6.4 illustrates the change of λ^* and λ_t with respect to patient e-visit usage α. Let α^* be the optimal diversion, when $\alpha < \alpha^*$, e-visit is underutilized, and the maximum arrival rate λ^* is increasing with respect to α. When $\alpha > \alpha^*$, the maximum arrival rate λ^* decreases. By maximizing the value of $\frac{\lambda^*}{\lambda_t}$ in (6.22), we have Corollary 6.3.

Corollary 6.3 *Under the assumption that the referral ratio β_{ov} is an linear function of the patient diversion factor α, i.e., $\beta_{ov} = k_1\alpha + k_2$, the optimal α^* to maximize λ^* in (6.22) is*

$$\alpha^* = \frac{1 - \gamma - k_2}{2k_1}.$$

Proof: See the Appendix. ■

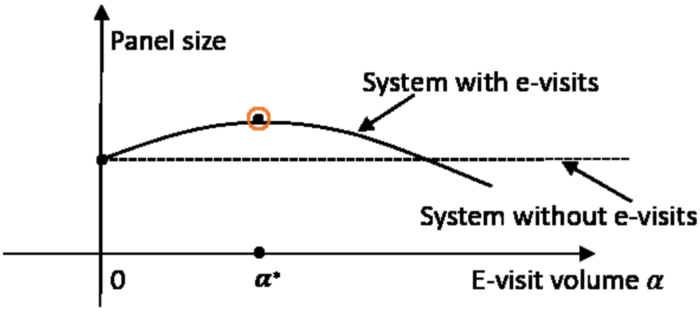


Figure 6.4: Panel size change with respect to patient diversion factor α

In close, if the referral ratio is not related to e-visit volume, then the larger the diversion, the more patients each physician would expect to handle. Now as β_{ov} is positively correlated with α, it is no longer the case that the more e-visits the better. The optimal diversion α^* is determined by the service rate ratio and how the referral ratio is affected by e-visit patient volume. As a remark, the relationship between α and β_{ov} might not be linear, and the values of k_1 and k_2 might not be easily obtained. The purpose of this discussion is to serve as a caveat to avoid abusing e-visits – physicians should advertise e-visit properly, and provide their patients with the most appropriate care.

6.4 Conclusions

In this chapter, formulas to evaluate the overall system performance of primary care delivery systems with and without e-visits are developed. The criteria for implementing e-visits are investigated. For system with e-visits to outperform the system without e-visits, conditions such as the referral ratio should be less than the relative service rate difference are required. In summary, to benefit from implementing e-visits, variables such as mean e-visit service time, e-visit service time variation, and the referral ratio from e-visits to office visit are all the smaller the better. Besides, adopting e-visits could potentially increase the physician's capacity to handle more patients. Physicians should also direct those patients who are suitable for e-visits to receive e-visits. In conclusion, this work provides a quantitative tool for primary care physicians to understand e-visit's impact on patients' accessibility to care.

Chapter 7

Case Studies

To elucidate the applicability of the methods introduced in this dissertation, we present two case studies to show how the modeling framework can be used to provide managerial insights and identify opportunities and challenges for future improvement. Specifically, in Section 7.1, the modeling approach is applied to a gastroenterology clinic for workflow redesign, and in Section 7.2, the mammography testing process is studied and the demand change analysis is conducted.

7.1 Case Study I: Design and Analysis of a Gastroenterology (GI) Clinic

The University of Wisconsin Medical Foundation was designing a new Digestive Health Center in Madison, Wisconsin. The DHC is a multi-disciplinary health care facility that integrates various services including gastroenterology and hepatology clinic, colorectal surgery clinic, endoscopy procedures, radiology, laboratory, and pharmacy. The DHC provides multi-disciplinary patient care and comprehensive clinical services related to the digestive tract, with specialists diagnosing and managing complex and chronic gastrointestinal disorders. The mission of the DHC is to partner with patients and families to meet the unique digestive health needs of every patient through comprehensive, unparalleled care while advancing research and educating the next generation of health professionals.

As an important part of the DHC, the GI clinic focuses on the digestive system and its disorders. Understanding the workflow in the current GI clinic can help streamline clinic operations and identify the opportunities for improvement in preparation for the new center. The goals of this work are to develop a quantitative model to analyze the patient flow in the new GI clinic, evaluate its design options, and propose recommendations for the ideal staff coverage needed to accommodate the anticipated patient volume.

7.1.1 Work Flow Description

The current GI clinic care provider team consists of a clinician (physician, physician assistant, or nurse practitioner) and one clinical staff (medical assistant or registered nurse). Two exam rooms are assigned to each provider team. The service times of the team members are random, but the variances of the service times are relatively small. Within the GI clinic, patient visits primarily fall into two categories: office visit (OFV) and consult visit (CON). The OFV is for patients who have frequent visits to a GI specialist due to a chronic GI illness requiring frequent clinician care. The visit type CON is for patients who are new to the GI service or recently confront an illness, often referred by other physicians (most frequently primary care physicians). Consult visits are scheduled for a longer duration than office visits. The office visits and consult visits are distributed throughout the daily schedule based on demand, provider preference, and office efficiency.

A typical visit to the GI clinic contains the following steps (see Figure 7.1):

- A patient checks in at the reception desk; a receptionist notifies the clinical staff, and the patient is seated in the waiting room.

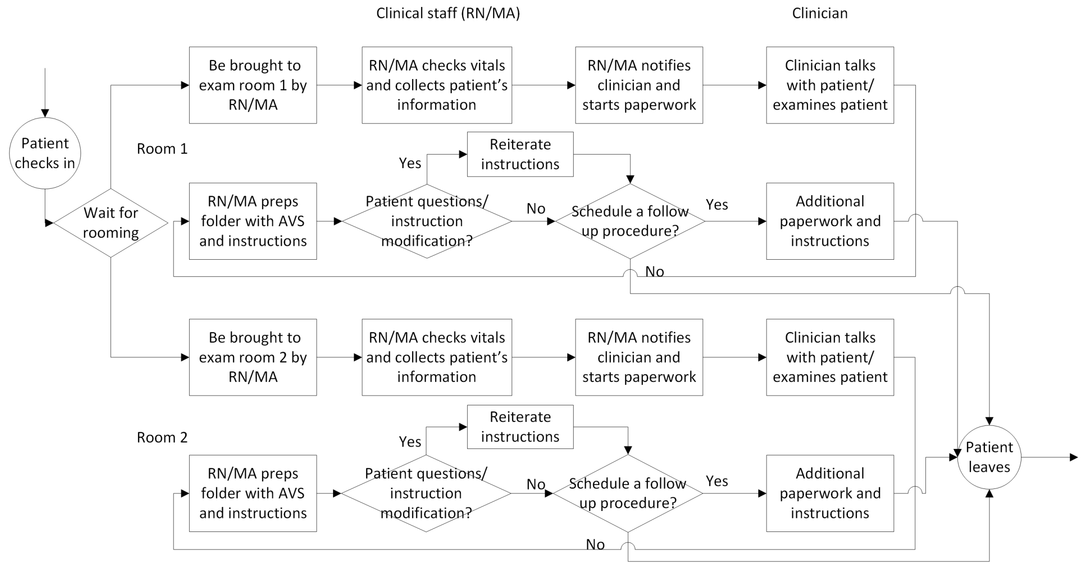

Clinical staff (RN/MA)
Clinician
Patient checks in
Wait for rooming
Room 1
Be brought to exam room 1 by RN/MA
RN/MA checks vitals and collects patient's information
RN/MA notifies clinician and starts paperwork
Clinician talks with patient/ examines patient
RN/MA preps folder with AVS and instructions
Patient questions/ instruction modification?
Yes
Reiterate instructions
No
Schedule a follow up procedure?
Yes
Additional paperwork and instructions
No
Be brought to exam room 2 by RN/MA
RN/MA checks vitals and collects patient's information
RN/MA notifies clinician and starts paperwork
Clinician talks with patient/ examines patient
Room 2
Reiterate instructions
Yes
RN/MA preps folder with AVS and instructions
Patient questions/ instruction modification?
No
Schedule a follow up procedure?
Yes
Additional paperwork and instructions
No
Patient leaves

Figure 7.1: Workflow in the GI clinic for one provider team

- The patient is escorted from the waiting room to an exam room by the clinical staff. The clinical staff collects basic information from the patient, obtains vitals, prepares paperwork, and records information into the health information systems. This step is referred to as patient rooming.

- The clinician enters the exam room, assesses and diagnoses the patient's condition, and develops a treatment plan.

- To discharge the patient, the clinical staff prepares the after visit summary (AVS) and follow-up instructions. The clinical staff then explains next steps and instructions to the patient and schedules any future clinic visits or procedures including colonoscopy, endoscopy, MRI and CT Scan, etc. If any appointments regarding the above procedures are needed, additional documentation is required after the appointment.

- Finally, the patient leaves the clinic.

7.1.2 System Modeling and Performance Analysis

Since each provider team works independently, we focus our study on one provider team primarily. In this case, the workflow can be simplified into a serial process which comprises of six steps: patient waiting for rooming, rooming (information collection, vital check, paperwork, reporting, etc.), patient waiting for the clinician, clinician examination and diagnosis, patient waiting for check-out, and check-out (including schedule possible follow-up appointments, file additional paperwork and give instructions, etc.). Finally, the patient leaves the clinic. Such a workflow is illustrated in Figure 7.2, where the circles represent the services, and the rectangles characterize patient waiting for the next service.

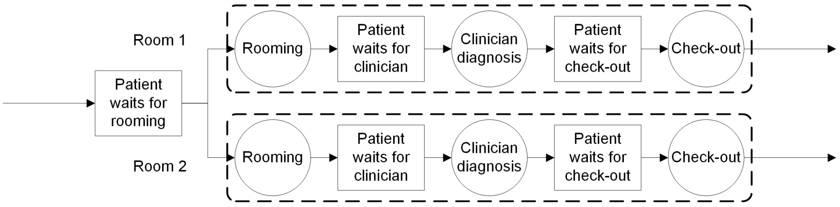

Room 1
Patient waits for rooming
Room 2
Rooming
Patient waits for clinician
Clinician diagnosis
Patient waits for check-out
Check-out
Rooming
Patient waits for clinician
Clinician diagnosis
Patient waits for check-out
Check-out

Figure 7.2: Structural model of the GI clinic workflow

To analyze the above process, the following assumptions and notations are introduced to address the services, the resources (clinicians and clinical staffs), and their interactions.

(i) For patient arrival, the inter-arrival time of the incoming patients follows the exponential distribution with arrival rate λ.

(ii) There are N steps after a patient arrives, where $N = 6$ in the current model (see the following list of processes). It is assumed that all six steps are identical for each room. There are three provider services in each exam room: clinical staff rooming, clinician visit, and clinical staff wrap-up, denoted as services 1, 2, and 3, respectively. For other waiting steps, the mean cycle times are all zeros. Here we assume the service times follow exponential distributions with corresponding processing rates $c_i = \frac{1}{\tau_i}$, $i = 1, 2, 3$. The extension to non-Markovian arrival and service will be estimated using the empirical formulas described in Chapter 4.

(a) patient waiting for rooming,

(b) patient rooming/clinical staff visit,

(c) patient waiting for clinician service,

(d) clinician examination and diagnosis,

(e) patient waiting for check-out,

(f) clinical staff checking out the patient.

(iii) The number of rooms assigned to one provider is M, where $M = 2$ in the current model. If a patient arrives while all the exam rooms are occupied, he/she needs to wait in the lobby. The maximum capacity of the waiting lobby is set as Q. In this study, we select $Q = 10$ according to the capacity of the clinic under study.

(iv) There are two types of resources in the system ($R = 2$). The number of resources is defined by $\{r_1, r_2\}$, representing the number of clinical staff and clinicians, respectively. In the current setting, $r_1 = r_2 = 1$.

(v) The staff allocation for each process is denoted as θ_i, $i = 1, 2, \ldots, 6$. The current configuration is $\{\theta_1, \ldots, \theta_6\} = \{0, 1, 0, 2, 0, 1\}$, where 0 implies that no resource is needed, and $\theta_2 = \theta_6 = 1$ and $\theta_4 = 2$ represent that the required resources for the second and sixth steps, and the fourth step are clinical staff and clinician, respectively.

(vi) Sometimes two services may require the same type of resource. In this case, priority is assigned to a later service. For example, if a patient needs to be discharged and another patient is waiting for rooming, the clinical staff will discharge the first patient and then room the other. There is no interruption of the ongoing service, i.e., if the resource is being used, the next patient has to wait until the current service finishes.

The detailed description of the transitions can be found in the Appendix and Zhong et al. [132]. With the balance equations (A.11), and by applying the same derivation scheme as in Section 3.3.4, the patient average length of visit and staff utilization can be obtained. Define TP as the system throughput rate, i.e., the rate patient leaving from the last service, and WIP as the average number of patients in the system. Then we obtain Theorem 7.1.

Theorem 7.1 *Under assumptions (i)-(vi), TP and WIP can be calculated as follows:*

$$TP = c_3 \sum_{l=1}^{K} P_l s_6^l, \tag{7.1}$$

$$WIP = \sum_{l=1}^{K} \left(P_l \sum_{j=1}^{6} s_j^l \right), \tag{7.2}$$

where $P_l's$ are solved using the same method as in (3.16 - 3.18).

By Little's Law, the patient average length of visit, T_s, can be obtained.

Corollary 7.1 *Under assumptions (i)-(vi),*

$$T_s = \frac{WIP}{TP} = \frac{\sum_{l=1}^{K} \left(P_l \sum_{j=1}^{6} s_j^l \right)}{c_3 \sum_{l=1}^{K} P_l s_6^l}. \tag{7.3}$$

In addition to patient length of visit, the staff utilizations can be calculated as follows:

Corollary 7.2 *Under assumptions (i)-(vi),*

$$\rho_{\textit{clinical staff}} = \sum_{l=1}^{K} P_l (s_2^l + s_6^l), \tag{7.4}$$

$$\rho_{\textit{clinician}} = \sum_{l=1}^{K} P_l s_4^l. \tag{7.5}$$

Remark 7.1 The length of visit in Corollary 7.1 is obtained under the assumption of exponential inter-arrival and service times, and is nominated as T_s^{exp}. When extending to non-Markovian scenarios, the length of visit can be estimated using the empirical formulas introduced in Subsection 4.3.3 based on T_s^{exp} and the corresponding distribution parameters.

7.1.3 Design of the New GI Clinic

The DHC will consolidate several satellite clinics and endoscopy locations into a single center and is expected to accommodate 14,500 GI clinic visits per year. They spare 15 exam rooms in the new center for GI services. The collaborative multi-disciplinary team consists of about 30 clinicians. However, not all rooms are open every day and not all physicians show up every day. The clinic will keep a schedule regarding which room is available and which clinician is at service each day. The 15 exam rooms are divided into independent pods and each clinician is working independently with two assigned exam rooms. To achieve a better service in the new GI clinic, the impacts of staffing alternatives, the number of rooms, and demand changes need to be apprehended. In addition, a new clinic layout and operational processes have been proposed. In particular, different check-out processes have been designed to improve patient access by decreasing the patient length of visit. Table 7.1 summarizes all the scenarios in the what-if analyses.

Note that the 50% clinical staff availability is intended to model the scenario where

Table 7.1: Summary of the what-if scenarios for designing the new GI clinic

Scenario	Category	Description
1	Staffing model	50% clinical staff availability or two clinical staffs per clinician
2	Demand change	Increase demand by 10% or 30%
3	Room configuration	One or three exam rooms per clinician
4	Service times	Change service times of clinical staff or clinician by 10%
5	Combined scenarios	Add one room or one clinical staff and increase demand by 30%

one clinical staff supports two clinicians so that roughly 50% of the clinical staff's effort is devoted to each clinician. In this case, the model discussed in Section 7.1.2 is still applicable with the modification that the rate of rooming and checking out the patient should be decreased by half approximately, i.e., $c_1' = \frac{c_1}{2}$, and $c_3' = \frac{c_3}{2}$. This implies that the patient may stay at the previous state after finishing it, due to the unavailability of the clinical staff. Finally, the last scenario is a combination of all parameter changes in scenarios 1-3. In the following subsections, the performance evaluations of these scenarios are introduced.

Staffing model

First, we investigate the impact of changes in the current staffing model. Instead of having one clinical staff to assist one clinician, we inspect the case of one clinical staff supporting two clinicians (i.e., 50% clinical staff availability for each clinician), and the case of two clinical staff for each clinician. The results are summarized in Table 7.2.

The above results manifest that a clinical staff of 50% availability is definitely not enough since the patient average length of visit increases significantly and the utilization of the clinical staff is doubled. However, the case of two clinical staff for one clinician is not necessary since it significantly decreases clinical staff's utilization while the decrease in the average length of visit is not remarkable. Therefore, the current staffing model of one clinical staff for one clinician can well accommodate the current demand.

Remark 7.2 Note that the staff utilization obtained from the model only represents the time percentage the clinician or clinical staff is working with the patient inside the exam room. In addition to serving patients (rooming, diagnosis, medication, etc.), they also carry out a substantial amount of work outside the patient rooms, such as documenting/reporting, answering phone calls/messages, and analysis of lab testing results.

Table 7.2: GI clinic: staffing model comparison

(a) 50% clinical staff availability per clinician

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	117.7	117.32
$\rho_{\text{clinical staff}}(\%)$	43.95	81.14	84.61
$\rho_{\text{clinician}}(\%)$	47.06	43.44	-7.69

(b) Two clinical staffs per clinician

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	48.22	-10.9
$\rho_{\text{clinical staff}}(\%)$	43.95	21.99	-49.97
$\rho_{\text{clinician}}(\%)$	47.06	47.09	0.06

It is assumed that the service in the exam room has a higher priority so that the other activities will be stopped whenever a patient service is requested.

Demand change

Next, we study the effects of patient demand change on system performance. The current inter-arrival times of 15 and 45 minutes are dictated by the clinic scheduling system. We investigate the system with the same structural model, but with decreased inter-arrival times (for instance, from 15 to 13.5 minutes, and 45 to 40.5 minutes, for a 10% increase in demand; and to 10.5 and 34.6 minutes, for a 30% increase in demand).

Remark 7.3 In real practice, when patient demand for the care provider increases, instead of changing the scheduling template to make shorter slots, which might yields messy schedules, the scheduler tends to double book or triple book patients into the original slots. However, because of the variability in arrival, directly decreasing inter-arrival times would yield similar patient arrival patterns.

Table 7.3: GI clinic: patient demand change

(a) Demand increased by 10%

	From	To	Changes (%)
LOS_{model} (min)	54.16	58.11	7.29
$\rho_{\text{clinical staff}}(\%)$	43.95	48.14	9.53
$\rho_{\text{clinician}}(\%)$	47.06	51.54	9.52

(b) Demand increased by 30%

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	70.32	29.85
$\rho_{\text{clinical staff}}(\%)$	43.95	56.06	27.56
$\rho_{\text{clinician}}(\%)$	47.06	60.02	27.54

As advertised in Table 7.3, if the demand is increased by 10%, the increase in the average length of visit is 7.29%, which is not favorable, but still can be accommodated. However, the current GI Clinic does not have the capability to comply a 30% demand surge – the average length of visit increases substantially under this setting. In addition, both clinical staff and clinician utilizations are increased by about 30%. Although more patients can be served, the excessive waiting time for the patients and substantial overtime work for the providers are ineluctable. More capacity and resources are demanded in this scenario.

Room configuration

Table 7.4: GI clinic: room configuration change

(a) One exam room

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	81.92	51.26
$\rho_{\text{clinical staff}}(\%)$	43.95	42.77	-2.68
$\rho_{\text{clinician}}(\%)$	47.06	45.79	-2.77

(b) Three exam rooms

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	51.22	-5.43
$\rho_{\text{clinical staff}}(\%)$	43.95	43.99	0.09
$\rho_{\text{clinician}}(\%)$	47.06	47.11	0.11

Here we change the number of rooms assigned to each provider group to one and three. The results are compared in Table 7.4. On one hand, dropping one room increases the patient average length of visit by 51.26%, which indicates that one room is not enough and causes a long wait for rooming. On the other hand, by adding one more room, the length of visit is decreased by 5.43%, which is not significant. Therefore, the current setting of two rooms per provider team is reasonable.

Service times

The change in service times of both clinical staff and clinician are investigated. Suppose the service times of the clinical staff and the clinician are decreased by 10%. The corresponding performance is presented in Table 7.5.

Table 7.5: GI clinic: service time change

(a) Decrease clinical staff service times by 10%

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	51.66	-4.62
$\rho_{\text{clinical staff}}(\%)$	43.95	39.57	-9.97
$\rho_{\text{clinician}}(\%)$	47.06	47.08	0.04

(b) Decrease clinician service time by 10%

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	51.17	-5.52
$\rho_{\text{clinical staff}}(\%)$	43.95	43.98	0.07
$\rho_{\text{clinician}}(\%)$	47.06	42.38	-9.95

As one can see, decreasing the service time of either the clinical staff or the clinician would have the similar impact on system performance, due to their similar workloads in the current system setting. Usually, the patient average length of visit is more sensitive to the service with a longer operation time, which is clinician's service in this system. From our observation, the clinician and the clinical staff may ask the patient the same questions repeatedly during their visits. Therefore, improving coordination between the clinician and the clinical staff to decrease duplicate work could possibly result in reduced staff service time. Additionally, some of the paperwork can be prepared by the clinical staff during the clinician's visit so that the patient check-out time could be reduced. Furthermore, some information for patients with frequent visits can be prepared prior to the visit. Thus, there exist considerable opportunities to reduce service time without sacrificing care quality and patient satisfaction.

Combined Scenarios

Finally, we study the scenario that multiple parameters are subject to change. In this circumstance, the demand is increased by 30%, and at the same time, one more room is added, or one more clinical staff is added to the system.

Table 7.6: GI clinic: combined scenarios

(a) Increase demand by 30% and add a room

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	61.26	13.12
$\rho_{\text{clinical staff}}(\%)$	43.95	56.89	29.46
$\rho_{\text{clinician}}(\%)$	47.06	60.86	29.32

(b) Increase demand by 30% and add a clinical staff

	From	To	Changes (%)
$LOS_{model}(min)$	54.16	54.02	-0.26
$\rho_{\text{clinical staff}}(\%)$	43.95	25.65	-41.64
$\rho_{\text{clinician}}(\%)$	47.06	61.04	29.71

When only demand is increased (see Table 7.3), a 30% demand surge leads to a roughly 30% increase in the average length of visit. However, such an increase shrinks to 13% when an additional room is introduced (Table 7.6). On the contrary, if an additional clinical staff is added, even with 30% demand increase, the average length of visit will not increase, but decrease by 0.26%. Therefore, additional clinical staff would be needed to reconcile the high volume of patients.

In summary, the above what-if analyses pinpoint the desired clinical setting: the current setting of two exam rooms, one clinician, one clinical staff could efficiently accommodate the current patient demand. A 10% increase in patient demand can be accommodated with a degraded performance, but a 30% increase will need an extra clinical staff to ensure the desired quality of care. Reducing the staff service time could be beneficial, but the implementation needs to be further investigated so that patient outcomes will not be sacrificed. Moreover, it can be observed that the patient average length of visit is monotonically non-increasing with respect to the numbers of rooms and staff. These results offer a quantitative guideline for designing the new GI clinic.

Check-Out Process

Three check-out processes are considered:

(I) *Check-out in exam room*: In this process, the patient would check out in the same room where he/she is examined. Such a model has been studied in previous subsections.

(II) *Check-out in scheduling room*: This process suggests using a dedicated scheduling room and scheduler to perform the check-out function for every four exam rooms. Therefore, check-out does not occur in the exam room. The idea of this process is to increase the availability of exam rooms to incoming patients.

(III) *Mixed check-out process*: In this case, the patients can either check out inside the exam room or go to the scheduling room shared by the four exam rooms.

To investigate the impact of different check-out processes, the analytical model described above has been modified to fit into each scenario, and the corresponding system performances are compared.

First, assume the check-out times are the same either inside the exam room or in the scheduling room. Since only one provider team is considered in the model, we have $c_4 = \frac{c_3}{2}$. Then, we evaluate the patient average length of visit as a function of the probability of check-out in the exam room, p, and the check-out service time τ_4. The results are exhibited in Figure 7.3, in which four check-out times, long, medium long, medium short, and short, denoted as $\tau_{4,j}$, $j = 1, \dots, 4$, are considered. In addition, we assume $\tau_{4,j} > \tau_{4,j+1}$, $j = 1, 2, 3$. Then, the following observations are obtained:

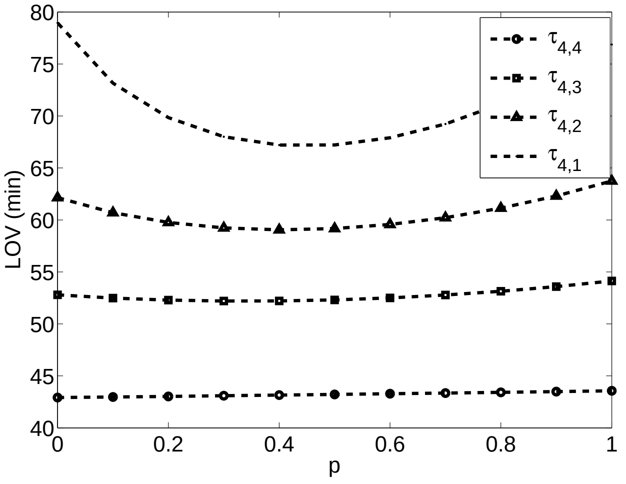


Figure 7.3: GI clinic check-out process comparison: identical check-out time

- When the check-out time is short, having most of the patients checking out in the scheduling room leads to a shorter average length of visit. This is due to an extra resource (scheduler) is added to the system, which is similar to adding a clinical staff with 50% availability in the original system.

- When the check-out time is long, a higher percentage of patient checking out inside the exam room will result in a better outcome, since the scheduler will take doubled workload (due to supporting two providers) and the check-out process in the scheduling room is even slower.

- In most cases, the mixed check-out process has the best performance. This is because the discharge workload assigned to the clinical staff and the scheduler are more or less balanced. These results imply that an effort to balance the workload will be beneficial in reducing the length of visit.

- When the check-out time is long, different check-out processes may result in a significant difference in patient length of visit. As revealed in Figure 7.3, the mixed check-out process could reduce almost 15% of the average length of visit. These results suggest that a proper check-out process could decrease the patient length of visit without utilizing extra workforce.

Next, assume the scheduler has a different discharge service time comparing with that of the clinical staff. This experiment is motivated by the possible scenario that the service time of the scheduler and the clinical staff can be significantly different, which might affect the choice of check-out process. In this case, introduce parameter α which characterizes the ratio between the two service times, i.e., $c_4 = \alpha c_3$ (or $\tau_4 = \frac{1}{\alpha}\tau_3$). Since the scheduler supports two provider teams, the service time is doubled looking from either one team's perspective. Thus, $\alpha = 0.5$ implies the two service times are identical.

In Figure 7.4, the results for three cases, $p = 0.25$, 0.5, and 0.75, are compared. By examining the figure, the following observations can be obtained:

- It should be pointed out that the patient average length of visit is monotonically decreasing with respect to α. Such monotonicity is due to the fact that a larger

α implies a shorter service time in the scheduling room. When α is large enough, the check-out time becomes extremely short and its impact on the overall length of visit becomes negligible. Consequently, we observe that when α is small, the patient average length of visit decreases fast, while when α is close to or larger than one, the decrease becomes minimal.

- When α is smaller than 0.4 (i.e., check-out in the scheduling room is slower than that in the exam room), the larger the probability p, the shorter the length of visit. In this case, more people should stay in the exam room to check out. When α is larger than 0.5 (check-out in the exam room takes more time), more people should check out in the scheduling room since now a smaller p leads to a shorter length of visit.

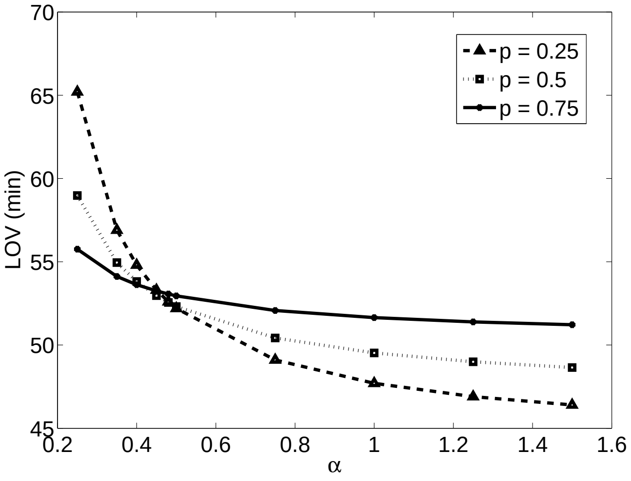


Figure 7.4: GI clinic check-out process comparison: different check-out times

From the above discussions, based on the configuration plan of the designed GI clinic, the mixed check-out process is recommended. If the check-out service time is long, more patients should check out inside the exam rooms. Otherwise, more patients are recommended to check out in the scheduling room, which would balance the discharge workload of the clinical staff and the scheduler, and, therefore, lead to a shorter patient length of visit on average. The UWMF leadership team has accepted the recommendation. It was applied in the final design and operation of the new GI clinic.

7.1.4 Conclusions

In this section, the analytical framework proposed in Chapter 3 is utilized to analyze the design of the workflow in a gastroenterology clinic. The patient average length of visit and the staff utilization are evaluated. What-if analyses are carried out to investigate the impacts of different workforce and resource configurations. It demonstrates that the allocation of one clinical staff and two exam rooms per clinician can well accommodate the current patient demand. If patient demand increases, adding a clinical staff is more effective to maintain the system performance than adding exam rooms. In addition, different check-out processes have been compared. The results reveal that the mixed check-out process that patients can check out either inside the exam room or in the scheduling room is the optimal way to ensure a smooth workflow and enhanced patient outcomes. Our model helps the leadership team understand the workflow in hospital units and clinics, streamline service operations, and identify the opportunities for improvement in preparation for the redesign.

7.2 Case Study II: Mammography Testing Process Modeling

With the rapid growth in health service demand, the efficient and safe use of radiology services for diagnosis and treatment is of extreme importance for the well-being of both patients and care providers. Mammography, which uses low-energy X-rays and allows the visualization of fine details in the breast tissue, is regarded as the most effective tool for routine breast cancer screening and early diagnose of cancer [133]. To ensure the effective use of mammography, the patient-flow analysis and work management are craved.

A case study of the mammography testing process patient flow at the Breast Imaging Center of the University of Wisconsin Medical Foundation in Madison, Wisconsin is introduced. In recent years, the imaging center of the UWMF has experienced an increasing demand for mammography testing for the detection of breast cancer. In 2012, the imaging center conducted around 11,000 mammography procedures, including 7,400 screening mammograms and 3,600 diagnostic mammograms, and thousands of breast ultrasounds, bone densities, biopsies, and breast MRIs. From 2013, the clinic plans to collaborate with the University of Wisconsin Carbone Cancer Center, which will bring an estimated 1,000 influx of mammography patients into the Breast Imaging Center. Therefore, the goal of this study is to develop a quantitative model of the mammography patient flow, investigate the impact of demand change, and propose recommendations.

7.2.1 Process Description and Structural Modeling

A typical Breast Imaging Center consists of mammography equipment, exam rooms, receptionists, technologist assistants (TA), radiology technologists (Tech), and imaging radiologists. The capacity of these resources varies according to test center sizes, demands, and purposes. However, the general procedures are usually standardized. Screening and diagnostic imaging are performed at designated exam rooms. Usually, the receptionist deals with all types of patient visits at the reception desk. The TA is responsible for bringing patients from the reception area to the changing room, preparing patient paperwork, dealing with schedule changes, and all other miscellaneous work. The Tech's are usually dedicated to their specific exam rooms taking images. An imaging radiologist is required when working with a diagnostic patient. The radiologist is not dedicated to the mammography unit but also working for other radiology departments and is seldom considered as a constraint in this system. Thus, in this study, we view the TA, Tech, and the exam room (equipment) as the primary constraints of interest. The workflow of a typical Breast Imaging Center is displayed in Figure 7.5.

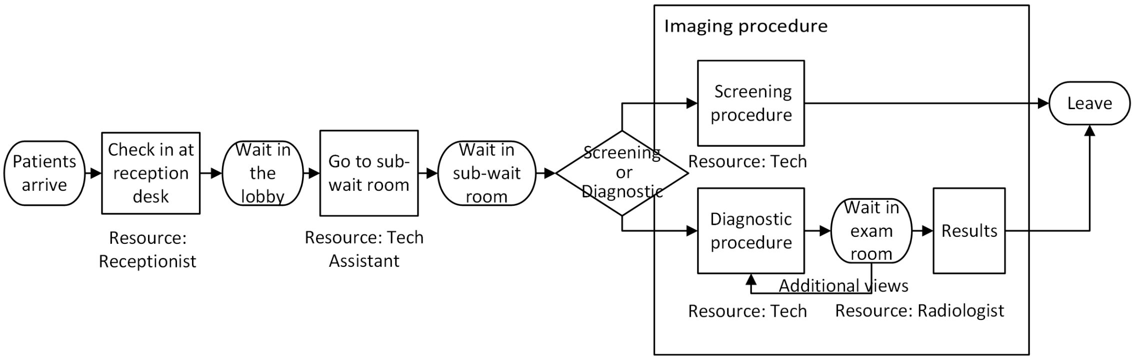


Figure 7.5: Mammography patient flow in the Breast Imaging Center

Each technologist works independently in their exam rooms. Therefore, we focus our study on one exam room initially and extend to the system with more exam rooms later.

In this case, the workflow can be simplified into a serial process which includes patient check-in (filling in forms), rooming by TA (changing into gown), imaging procedure (for screening patients, this procedure only involves image taking, while for diagnostic patients, radiologist reviews are included), and finally, patient changing clothes and leaving. Such a workflow is illustrated in Figure 7.6.

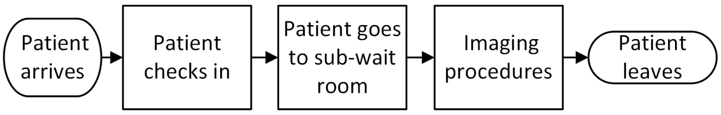


Figure 7.6: Mammography patient flow model: one exam room

Within each exam room, only one patient is permitted at a time. However, the TA will bring the next patient into a sub-waiting room for preparation right before the patient's scheduled imaging time, even if the exam room is still occupied. So there is a maximum of two patients in the sub-waiting room and exam room. Typically, the TA leaves once the patient finishes changing into a gown. Moreover, since most of the patients arrive ahead of their appointment time, a finite capacity of the waiting area is assumed for these patients. However, the number of patients waiting for the exam room will not be too large, due to the fact that the appointments are scheduled in advance.

For the imaging procedure, diagnostic patients have to wait for results and comments from the radiologist which demands a longer stay, while for screening patients, the stay is relatively shorter. Thus, the service times of the radiology technologist among different types of patients can be significantly different. Except this, the other service times for different types of visits are relatively similar. Note that although the technologists work independently, all the exam rooms share only one TA, which may introduce an availability issue and cause delay.

7.2.2 Model Development and Validation

The Breast Imaging Center at the UWMF performs mainly screening, diagnostic imaging, and bone density tests. The center consists of a changing (sub-waiting) room, a bone density room, an ultrasound room, a staff lounge, and three examination rooms, each equips with one mammography machine. Due to the high demand in screening, two of the exam rooms are used for screening test while the third is for diagnostic imaging. There are about 45-60 daily visits for screening and diagnostic mammography in total. The care provider team consists of three radiology technologists, two imaging radiologists, and one technologist assistant. The technologists are cross-functional so that they can perform both diagnostic and screening imaging. However, they are dedicated to one exam room per each shift. The TA is responsible for bringing patients from the reception area (waiting room) to the changing room (sub-waiting room), preparing patient paperwork, dealing with schedule changes, and any other miscellaneous work. A diagnostic appointment is scheduled for every 30 minutes while a screening appointment is for every 20 minutes, starting from eight in the morning. Such a workflow is illustrated in Figure 7.7.

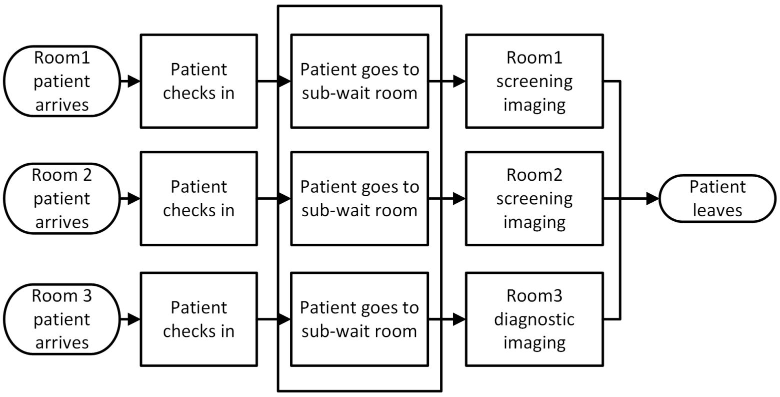


Figure 7.7: UWMF Breast Imaging Center patient flow model

Using the method introduced in Chapter 4, an analytical model has been developed to characterize the Breast Imaging Center. To validate the model, the results obtained from the model analysis are compared with that obtained from hundreds of observations and records from heath information system database within one month in the Breast Imaging Center. Same simulation setups are used for validation purposes and the confidence intervals are typically within 2% of the performance measure.

Let $LOS_{observed}$ and LOS_{model} denote the average lengths of visit obtained by data collection and the analytical model, respectively. Introduce

$$\begin{aligned} \triangle &= LOS_{observed} - LOS_{model}, \\ \epsilon &= \frac{LOS_{observed} - LOS_{model}}{LOS_{observed}} \cdot 100\%. \end{aligned}$$

The results of such comparisons are summarized in Table 7.7. As one can see, the differences between them are minor. Therefore, the model can accurately estimate the system performance and is proper for carrying out further analysis.

Table 7.7: Mammography testing process model validation

	Screening	Diagnostic
$LOS_{observed}$ (min)	25	34
LOS_{model} (min)	25.2	32.9
$\triangle$ (min)	-0.2	1.1
ϵ	-0.8%	3.24%

7.2.3 Demand Change Analysis

Using the model, we investigate the impact of patient demand change. To respond to a demand increase, without loss of generality, the scheduled inter-arrival times are decreased. With a 5% demand increase, the inter-arrival times are decreased from 20 to 19 minutes and from 30 to 28.5 minutes, for screening and diagnostic patients, respectively. For a 10% increase in demand, such times are reduced to 18 and 27 minutes, respectively. Note that with the increasing demand, there is a higher possibility that more patients will wait in the queue. To avoid the scenario that patient will be rejected in the analytical model due to limited queue length, the queue size for each room is increased to $Q_i = 10$ and $Q_i = 20$, $i = 1, 2, 3$, for 5% and 10% demand increase, respectively. The consequence changes in the patient length of visit and staff utilization in one room are illustrated in Table 7.8.

As one can see, a 5% demand increase will lead to an 18.3% increase in the length of visit for screening patients and a 13.7% increase for diagnostic patients. The utilization of the TA and the Tech of both patient types is increased by 6% and 9%, respectively. Although not favorable, such demand change can still be accommodated with the current clinic setting.

However, the Breast Imaging Center does not have the capacity to accommodate a 10% demand surge. In this scenario, the patient length of visit will increase substantially, with a 45% and 41% spike for screening and diagnostic patients, respectively. In addition, the utilization of the TA and the Tech is increased by 12% and 16%, respectively.

Although more patients can be served, the results of excessive waiting time and substantial overload for the providers are not desirable. More capacity and resources are demanded in this scenario (note that the provider utilization is for one room and only involves the work in contact with patients, while many other responsibilities are

Table 7.8: Mammography patient demand change

(a) Demand increased by 5%

Screening patient	From	To	Changes (%)
$LOS_{\text{screening}}$ (min)	25.2	29.8	18.3
$\rho_{\text{TA}}(\%)$	5.0	5.3	6.0
$\rho_{\text{tech}}(\%)$	70.8	76.7	8.3
Diagnostic patient	From	To	Changes (%)
$LOS_{\text{diagnostic}}$ (min)	32.9	37.4	13.7
$\rho_{\text{TA}}(\%)$	3.3	3.5	6.1
$\rho_{\text{tech}}(\%)$	77.6	84.5	8.9

(b) Demand increased by 10%

Screening patient	From	To	Changes (%)
$LOS_{\text{screening}}$ (min)	25.2	36.6	45.2
$\rho_{\text{TA}}(\%)$	5.0	5.6	12
$\rho_{\text{tech}}(\%)$	70.8	82.0	15.8
Diagnostic patient	From	To	Changes (%)
$LOS_{\text{diagnostic}}$ (min)	32.9	46.6	41.6
$\rho_{\text{TA}}(\%)$	3.3	3.7	12.1
$\rho_{\text{tech}}(\%)$	77.6	90.7	16.9

not included).

To accommodate such demand changes, several possible solutions are proposed, which include extending work time (either starting work earlier or finishing later or shrinking break time), diffusing the patients to other clinic sites, or adding extra exam rooms with equipment and technologists. Since the TA's workload in direct contact with patients is not high, there is no need to increase the number of TAs. In this case, we test the scenario that one additional exam room and one more Tech are added. With the same amount of arrival, the new arrival rate is decreased by 25% for each exam room (due to adding one room). We compare the two scenarios (3 exam rooms, 3 technologists, and 4 exam rooms, 4 technologists) in respect to the average patient length of visit.

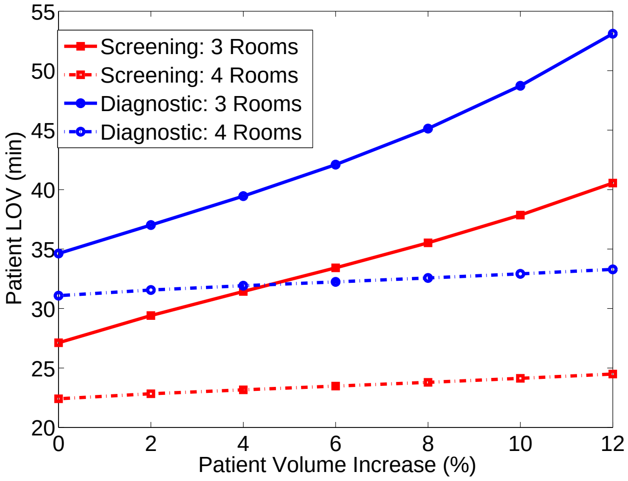


Figure 7.8: Mammography patient LOV change w.r.t. increasing demand

From the results exhibited in Figure 7.8, we can conclude that the current system is running at a relatively high intensity, i.e., the patient length of visit increases rapidly with respect to the demand increase. While in four exam rooms case, the workload of each room is reduced, and the increase of the patient length of visit is moderate with a higher demand. With as much as 12% increase in demand, the increase in screening patients and diagnostic patients' length of visit is only around 2 minutes. Therefore, with capacity (room, equipment, and technologist) increase, the Breast Imaging Center will be able to accommodate the surge in patient volumes. The recommendation has been acknowledged by the clinic leadership team.

7.2.4 Conclusions

In this section, an analytical model is developed and the iterative method is applied to study the workflow of mammography testing process. Using this model, demand change analysis is carried out and it manifests that with a substantial demand increase, extending the work time or adding more equipment and resource is demanded. In future work, using such a model, we can investigate the optimal control policies. For example, the model can be used to determine the minimum number of TAs required to achieve the desired patient length of visit.

Chapter 8

Summary and Future Research

8.1 Summary

Modeling care delivery services is critical to achieve efficient patient flow in hospital divisions and clinics. Health professionals are increasingly aware of the need to use their resources as efficiently as possible to empower proper care delivery. As the core part of my dissertation, a theoretical framework to characterize primary care delivery systems by modeling care providers' activities and patient flow is established. This is a challenging problem for two reasons. First, healthcare delivery systems can be intrinsically complicated featuring multiple care activities conducted within multiple patient rooms by a limited number of care providers. Secondly, models to characterize healthcare delivery systems need to be highly attuned to the subtleties of human behaviors and accommodate system variability. To tackle these challenges, innovative models and methods are proposed in this dissertation.

In particular, two ways to improve primary care delivery are investigated. To improve the efficiency of the current care delivery channel, stochastic models are developed to characterize care activities within patient rooms to streamline patient flow. It's demonstrated that with an appropriate level of simplification and abstraction, Markovian chain based models can provide a deft analysis of the care delivery processes while preserving the realism of the system. One novel work is the development of a convergent iterative method to tackle the problems of resource sharing and non-identical services. The proposed shared resource iteration overcomes the issues of the curse of dimensionality for scaled-up systems and makes it feasible to work with heterogeneous and interdependent systems to extend the modeling scope to the care facility level. In addition, this work also inspects electronic visit as a new channel for care delivery and investigates its impact, where few analytical study prevails. A queueing model is built to address physicians' operations on both office visits and e-visits and incorporate physicians' other tasks not in direct contact with patients. System-theoretic properties can be analyzed.

As this dissertation has attempted to demonstrate, effective analytical models can facilitate the improvement in patient's accessibility to care, enlighten the design of delivering the most appropriate care in a timely fashion, and enhance the provider's productivity.

8.2 Future Work

The future direction of health care delivery will be more intelligent, flexible and patient-centered. The emerging information technologies such as electronic visit, telemedicine, and teleconsult would bring drastic changes to the traditional care delivery system. It's demanding that innovative methodologies with various optimization approaches should be introduced to 1) integrate diverse care delivery options to investigate their interactions and impacts and 2) enhance the design of delivering the most appropriate care. Furthermore, customized care delivery is desired where care can be delivered timely and proactively, targeting on the right population. Along this line, it's necessary to embark on developing scalable algorithms to transform massive electronic medical record data into patient health information and integrating them with service system dynamics to accelerate smart care delivery.

Combining the data analytics with stochastic modeling, my future research aims to construct a two-level closed-loop hierarchical framework. At the lower-level, analytical modeling evokes and defines data, and in return, data analytics advances and refines analytical modeling. Such a framework will enable the optimal decision making and system control, which generates managerial insights and feedback to reinforce system modeling and data analytics at the upper-level. The broader impact lies in the discovery of the underlying laws that govern diverse sets of service systems sharing similar problem structures. An illustration of the framework is exhibited in Figure 8.1.

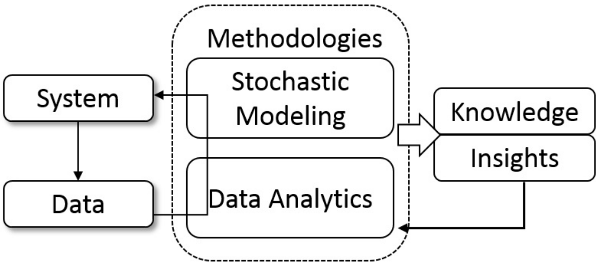


Figure 8.1: Stochastic modeling and data analytics framework

Appendix: Proofs and Derivations

Proofs of Chapter 3

Proof of Corollary 3.1: When $M = 2, N = 2$ and $R = 2$, there exist three states in total. Define the states as:

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

where the elements in S represent the number of patients in services 1 and 2. Since $r_1 = r_2 = 1$ (i.e., one resource in each service, such as nurse and doctor), we have

$$R = 2; \quad \mathcal{R} = [1,1]; \quad \Theta = [1,2].$$

Let P_i denote the probability the system is in state S_i, $i = 1,2,3$. Then the transition equations are obtained as follows:

$$\begin{aligned} c_1 P_1 &= c_2 P_2, \\ c_1 P_2 &= c_2 P_3. \end{aligned}$$

In addition,

$$P_1 + P_2 + P_3 = 1.$$

Solving P_i we obtain:

$$\begin{aligned} P_1 &= \frac{1}{1 + (\frac{c_1}{c_2}) + (\frac{c_1}{c_2})^2}, \\ P_2 &= \frac{(\frac{c_1}{c_2})}{1 + (\frac{c_1}{c_2}) + (\frac{c_1}{c_2})^2}, \\ P_2 &= \frac{(\frac{c_1}{c_2})^2}{1 + (\frac{c_1}{c_2}) + (\frac{c_1}{c_2})^2}. \end{aligned}$$

Therefore, the system throughput TP is

$$\begin{aligned} TP &= (P_2+P_3)c_2 \\ &= \frac{(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2}{1+(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2}c_2. \end{aligned}$$

Then the patient length of visit and resource utilizations are obtained:

$$\begin{aligned} T_s = \frac{2}{TP} &= \frac{2[1+(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2]}{[(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2]c_2} = \frac{2(c_1^2+c_2^2+c_1c_2)}{c_1c_2(c_1+c_2)} \\ &= \frac{2(\tau_1^2+\tau_2^2+\tau_1\tau_2)}{\tau_1+\tau_2}, \\ \rho_1 = P_1+P_2 &= \frac{1+(\frac{c_1}{c_2})}{1+(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2} = \frac{c_2(c_1+c_2)}{c_1^2+c_2^2+c_1c_2} \\ &= \frac{\tau_1(\tau_1+\tau_2)}{\tau_1^2+\tau_2^2+\tau_1\tau_2}, \\ \rho_2 = P_2+P_3 &= \frac{(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2}{1+(\frac{c_1}{c_2})+(\frac{c_1}{c_2})^2} = \frac{c_1(c_1+c_2)}{c_1^2+c_2^2+c_1c_2} \\ &= \frac{\tau_2(\tau_1+\tau_2)}{\tau_1^2+\tau_2^2+\tau_1\tau_2}. \end{aligned}$$

■

Proof of Corollary 3.2:

$$\begin{aligned} \frac{\partial T_s}{\partial \tau_1} &= \frac{4\tau_1+2\tau_2}{\tau_1+\tau_2} - \frac{2(\tau_1^2+\tau_2^2+\tau_1\tau_2)}{(\tau_1+\tau_2)^2} \\ &= \frac{2}{(\tau_1+\tau_2)^2}[(2\tau_1+\tau_2)(\tau_1+\tau_2)-(\tau_1^2+\tau_2^2+\tau_1\tau_2)] \\ &= \frac{2\tau_1(\tau_1+2\tau_2)}{(\tau_1+\tau_2)^2} > 0. \end{aligned}$$

Analogously,

$$\frac{\partial T_s}{\partial \tau_2} = \frac{2\tau_2(\tau_2+2\tau_1)}{(\tau_1+\tau_2)^2} > 0.$$

Therefore, the monotonicity with respect to τ_1 and τ_2 exist. ■

Proof of Corollary 3.3:

$$\begin{aligned}
\frac{\partial\rho_1}{\partial\tau_1} &= \frac{2\tau_1+\tau_2}{\tau_1^2+\tau_2^2+\tau_1\tau_2} - \frac{\tau_1(\tau_1+\tau_2)}{(\tau_1^2+\tau_2^2+\tau_1\tau_2)^2}(2\tau_1+\tau_2) \\
&= \frac{2\tau_1+\tau_2}{(\tau_1^2+\tau_2^2+\tau_1\tau_2)^2}[\tau_1^2+\tau_2^2+\tau_1\tau_2-\tau_1(\tau_1+\tau_2)] \\
&= \frac{2\tau_1+\tau_2}{(\tau_1^2+\tau_2^2+\tau_1\tau_2)^2}\tau_2^2 > 0.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{\partial\rho_1}{\partial\tau_2} &= \frac{\tau_1}{\tau_1^2+\tau_2^2+\tau_1\tau_2} - \frac{\tau_1(\tau_1+\tau_2)}{(\tau_1^2+\tau_2^2+\tau_1\tau_2)^2}(2\tau_2+\tau_1) \\
&= \frac{\tau_1}{(\tau_1^2+\tau_2^2+\tau_1\tau_2)^2}[\tau_1^2+\tau_2^2+\tau_1\tau_2-(2\tau_2+\tau_1)(\tau_1+\tau_2)] \\
&= \frac{-\tau_1(2\tau_1+\tau_2)}{(\tau_1^2+\tau_2^2+\tau_1\tau_2)^2} < 0.
\end{aligned}$$

Therefore, ρ_1 is monotonicity increasing and decreasing with respect to τ_1 and τ_2, respectively. Similar arguments can be applied to prove the monotonicity of ρ_2. ■

Proof of Corollary 3.4:

$$\frac{\partial T_s}{\partial\tau_1} - \frac{\partial T_s}{\partial\tau_2} = \frac{2\tau_1(\tau_1+2\tau_2)}{(\tau_1+\tau_2)^2} - \frac{2\tau_2(\tau_2+2\tau_1)}{(\tau_1+\tau_2)^2} = \frac{2(\tau_1-\tau_2)}{\tau_1+\tau_2}.$$

As one can see, the system is more sensitive to the service with the longer service time. Thus, the service with a larger τ_i is the impeding process (i.e., bottleneck). ■

Proof of Corollary 3.5: Let $\tau_1+\tau_2=\tau=\mathit{constant}$. Then

$$\begin{aligned}
T_s &= \frac{\tau_2(\tau_1^2+\tau_2^2+\tau_1\tau_2)}{\tau_1+\tau_2} \\
&= \frac{2}{\tau}[\tau_1^2+(\tau-\tau_1)^2+\tau_1(\tau-\tau_1)] \\
&= \frac{2}{\tau}[(\tau_1-\frac{\tau}{2})^2+\frac{3}{4}\tau^2].
\end{aligned}$$

As one can see, T_s reaches the minimum when $\tau_1 = \frac{\tau}{2}$. Thus, the optimal allocation is $\tau_1^* = \tau_2^*$. ■

Proof of Corollary 3.6: Similar to the proof of Corollary 3.1, due to joint service and the higher priority of doctor visit, there only exist two states:

$$S_1 = (2,0), \qquad S_2 = (1,1).$$

Again let P_i denote the probability the system is in state S_i, $i = 1, 2$. Then the following equations are obtained:

$$\begin{aligned} c_1 P_1 &= c_2 P_2, \\ P_1 + P_2 &= 1. \end{aligned}$$

Solving P_i we obtain:

$$\begin{aligned} P_1 &= \frac{c_2}{c_1 + c_2}, \\ P_2 &= \frac{c_1}{c_1 + c_2}. \end{aligned}$$

Therefore, the system throughput TP is

$$TP = P_2 c_2 = \frac{c_1 c_2}{c_1 + c_2}.$$

Then the patient length of visit and resource utilizations are obtained:

$$\begin{aligned} T_s &= \frac{2}{TP} = 2 \cdot \frac{\frac{1}{\tau_1} \cdot \frac{1}{\tau_2}}{\frac{1}{\tau_1} + \frac{1}{\tau_2}} = 2(\tau_1 + \tau_2), \\ \rho_1 &= P_1 + P_2 = 100\%, \\ \rho_2 &= P_2 = \frac{\frac{1}{\tau_1}}{\frac{1}{\tau_1} + \frac{1}{\tau_2}} = \frac{\tau_2}{\tau_1 + \tau_2}. \end{aligned}$$

■

Proofs of Chapter 4

To prove Proposition 4.1, the following lemmas are needed:

Lemma 8.1 *Under assumptions 1)-4) in Section 4.2.2, in Procedure 4.1, if* $p_1^{(k)} < p_1^{(k-1)}$*, then* $p_1^{(k+1)} < p_1^{(k)}$*, for* $k \geq 2$.

Lemma 8.2 *Under assumptions 1)-4) in Section 4.2.2, in Procedure 4.1, if* $p_2^{(k)} > p_2^{(k-1)}$*, then* $p_2^{(k+1)} > p_2^{(k)}$*, for* $k \geq 2$.

Proof of Lemma 8.1: When $p_1^{(k)} < p_1^{(k-1)}$, we obtain

$$c_{2,2}^{(k)} = c_{2,2}(1 - p_1^{(k)}) > c_{2,2}(1 - p_1^{(k-1)}) = c_{2,2}^{(k-1)}.$$

As T_i is monotonically decreasing with respect to $c_{j,i}$ ([134], [135]), it can be concluded that

$$T_2^{(k)} < T_2^{(k-1)}.$$

This leads to

$$p_2^{(k)} = \frac{1}{c_{2,2}T_2^{(k)}} > \frac{1}{c_{2,2}T_2^{(k-1)}} = p_2^{(k-1)},$$

which follows that

$$p_2^{(k)} > p_2^{(k-1)}.$$

Continue such arguments, we have

$$c_{2,1}^{(k+1)} = c_{2,1}(1 - p_2^{(k)}) < c_{2,1}(1 - p_2^{(k-1)}) = c_{2,1}^{(k)},$$

and

$$T_1^{(k+1)} > T_1^{(k)}.$$

Then it follows that

$$p_1^{(k+1)} = \frac{1}{c_{2,1}T_1^{(k+1)}} < \frac{1}{c_{2,1}T_1^{(k)}} = p_1^{(k)}.$$

■

Proof of Lemma 8.2: The proof is similar to that of Lemma 8.1. ■

Proof of Proposition 4.1: From $p_2^{(0)} = 0$, we have

$$\begin{aligned} c_{2,1}^{(1)} &= c_{2,1}, \\ T_1^{(1)} &= f_T(Q_1, c_{1,1}, c_{2,1}^{(1)}, c_{3,1}, c_{4,1}, \lambda_1), \\ p_1^{(1)} &= \frac{1}{c_{2,1}T_1^{(1)}}. \end{aligned}$$

It follows that

$$\begin{aligned} c_{2,2}^{(1)} &= c_{2,2}(1 - p_1^{(1)}), \\ T_2^{(1)} &= f_T(Q_2, c_{1,2}, c_{2,2}^{(1)}, c_{3,2}, c_{4,2}, \lambda_2), \end{aligned}$$

which leads to

$$p_2^{(1)} = \frac{1}{c_{2,2}T_2^{(1)}} > 0.$$

Thus, we obtain

$$c_{2,1}^{(2)} = c_{2,1}(1 - p_2^{(1)}) < c_{2,1}^{(1)},$$

which implies that

$$T_1^{(2)} > T_1^{(1)},$$

and

$$p_1^{(2)} = \frac{1}{c_{2,1}T_1^{(2)}} < p_1^{(1)}.$$

Continue such arguments we obtain

$$c_{2,2}^{(2)} > c_{2,2}^{(1)}, \quad T_2^{(2)} < T_2^{(1)},$$

which implies that

$$p_2^{(2)} > p_2^{(1)}.$$

From Lemmas 8.1 and 8.2, we conclude that p_1 is monotonically decreasing, while p_2 is monotonically increasing. Since p_1 and p_2 are probabilities bounded by 0 and 1, by the theorem for convergence of a monotone sequence of real numbers, we conclude that as $n \to \infty$, $\{p_i^{(n)}\}$, $i = 1, 2$, are convergent, i.e.,

$$\lim_{n\to\infty} p_i^{(n)} = p_i, \quad i = 1, 2.$$

Using the convergent p_i, a convergent value of $T_i^{(n)}$ can be determined. Thus, we have

$$\lim_{n\to\infty} T_i^{(n)} = T_i, \quad i = 1, 2.$$

The proof of the convergence of the shared resource iteration to a unique value is sketched as follows. In mathematics, the Banach fixed-point theorem (also known as the contraction mapping theorem or contraction mapping principle) is an important tool in the theory of metric spaces. The Banach fixed-point theorem guarantees the existence and uniqueness of fixed points of certain self-maps of metric spaces, and provides a constructive method to find those fixed points.

Let (X, d) be a metric space. Then a map $F : X \to X$ is called a contraction mapping on X if there exists $q \in [0, 1)$ such that

$$d(F(x), F(y)) \leq qd(x, y)$$

for all x, y in X.

Banach fixed point theorem states that let (X, d) be a non-empty complete metric space with a contraction mapping $F : X \to X$, then F admits a unique fixed-point x^* in X (i.e. $F(x^*) = x^*$). Furthermore, x^* can be found as follows: start with an arbitrary element x_0 in X and define a sequence x_n by $x_n = F(x_{n-1})$, then $x_n \to x^*$.

For the shared resource iteration, starting with any arbitrary $p_i^{(0)}$, $i = 1, 2$, each iteration generates a new set of $p_i^{(k)}$, $i = 1, 2$. Therefore, we can view each iteration as a function F such that

$$p_i^{(k+1)} = F(p_i^{(k)}),\ k = 1, 2, \dots$$

Choose the measurement d as the absolute difference between the probabilities obtained from two consecutive iterations. Then, it is of interest to prove that

$$|p_i^{(k+1)} - p_i^{(k)}| = |F(p_i^{(k)}) - F(p_i^{(k-1)})| \leq q|p_i^{(k)} - p_i^{(k-1)}|.$$

According to the iteration procedure,

$$|p_i^{(k+1)} - p_i^{(k)}| = \frac{1}{c_{2,i}}|\frac{1}{T_i^{(k)}} - \frac{1}{T_i^{(k-1)}}|,$$

where $c_{2,i}$ is the transition rate of the second service in the $i-th$ room which is conducted by the shared resource. $T_i^{(k)}$ is the $i - th$ room's cycle time in the $k - th$ iteration.

Since for each iteration, all other variables are consistent except for $c_{2,i}$ changes, we only consider the service with the shared resource. Such a subsystem can be approximated by an M/M/1/C queue where C is the maximum number of patients at that service stage and $C = 2$ according to the assumption described in Section 4.2.3. Then we define the cycle time for the service with the shared resource in room i at the $k - th$ iteration as $t_i^{(k)}$. For this subsystem, denote the server intensity and service rate as $\rho_i^{(k)}$ and $\mu_i^{(k)}$ in room i during the $k - th$ iteration. Note that $\mu_i^{(k)} = c_{2,i}^{(k)}$.

According to the cycle time of an M/M/1/C queue:

$$\frac{1}{t_i^{(k)}} = \frac{(1+\rho_i^{(k)})}{(1+2\rho_i^{(k)})}\mu_i^{(k)}.$$

For the absolute difference between any two iterations,

$$\begin{aligned}|\frac{1}{t_i^{(k)}} - \frac{1}{t_i^{(k-1)}}| &= |\frac{(1+\rho_i^{(k)})}{(1+2\rho_i^{(k)})}\mu_i^{(k)} - \frac{(1+\rho_i^{(k-1)})}{(1+2\rho_i^{(k-1)})}\mu_i^{(k-1)}| \\ &= |\mu_i^{(k)} - \mu_i^{(k-1)}|\frac{(2\rho_i^{(k)}(1+\rho_i^{(k-1)}) + \rho_i^{(k)}(1+2\rho_i^{(k-1)}))}{(1+2\rho_i^{(k)})(1+2\rho_i^{(k-1)})} \\ &= q_i^{(k)}|\mu_i^{(k)} - \mu_i^{(k-1)}|,\end{aligned}$$

where $q_i^{(k)} = \frac{(2\rho_i^{(k)}(1+\rho_i^{(k-1)})+\rho_i^{(k)}(1+2\rho_i^{(k-1)}))}{(1+2\rho_i^{(k)})(1+2\rho_i^{(k-1)})}$.

After simple algebra operations it can be proved that $q_i^{(k)} < 1$ for any $0 \leq \rho_i^{(k)}, \rho_i^{(k-1)} < 1$. Let q_i be a real number satisfying $q_i \in [0,1)$ and q_i is the upper bound of the sequence $q_i^{(k)}$. Therefore,

$$|\frac{1}{t_i^{(k)}} - \frac{1}{t_i^{(k-1)}}| \leq q_i|\mu_i^{(k)} - \mu_i^{(k-1)}|.$$

Next, consider the shared resource iteration, for the first room:

$$\begin{aligned}|p_1^{(k+1)} - p_1^{(k)}| &= \frac{1}{c_{2,1}}|\frac{1}{T_1^{(k+1)}} - \frac{1}{T_1^{(k)}}| \\ &\leq \frac{1}{c_{2,1}}|\frac{1}{t_1^{(k+1)}} - \frac{1}{t_1^{(k)}}| \\ &\leq \frac{1}{c_{2,1}}q_1|\mu_1^{(k+1)} - \mu_1^{(k)}| \\ &= q_1|p_2^{(k)} - p_2^{(k-1)}|. \qquad \text{(A.1)}\end{aligned}$$

Similarly, for the second room:

$$\begin{aligned}|p_2^{(k+1)} - p_2^{(k)}| &= \frac{1}{c_{2,2}}|\frac{1}{T_2^{(k+1)}} - \frac{1}{T_2^{(k)}}| \\ &\leq \frac{1}{c_{2,2}}|\frac{1}{t_2^{(k+1)}} - \frac{1}{t_2^{(k)}}| \\ &\leq \frac{1}{c_{2,2}}q_2|\mu_2^{(k+1)} - \mu_2^{(k)}| \\ &= q_2|p_1^{(k+1)} - p_1^{(k)}|. \qquad \text{(A.2)}\end{aligned}$$

Combine the two scenarios together, it can be shown that

$$
\begin{aligned}
|p_1^{(k+1)} - p_1^{(k)}| \leq q_1|p_2^{(k)} - p_2^{(k-1)}| \leq q_1q_2|p_1^{(k)} - p_1^{(k-1)}|,\\
|p_2^{(k+1)} - p_2^{(k)}| \leq q_2|p_1^{(k+1)} - p_1^{(k)}| \leq q_1q_2|p_2^{(k)} - p_2^{(k-1)}|.
\end{aligned}
$$

In conclusion, $|p_i^{(k+1)} - p_i^{(k)}| \leq q|p_i^{(k)} - p_i^{(k-1)}|$, where $q = q_1q_2 < 1$, which satisfies the Banach fixed point theorem. Consequently, the shared resource iteration will converge to a unique set of probabilities p_i^*, $i = 1, 2$.

Remark 8.1 The first inequality in (A.1) and (A.2) is valid because T_i is the $i - th$ room's total cycle time but t_i only represents the cycle time for the service with the shared resource.

■

Proofs of Chapter 5

Due to space limitation, we omit the majority of algebraic operations and only provide the sketch of proofs.

Proof of Theorem 5.1: To model physician's office and e-visit services, consider an M/G/1 queue with vacations. Define N_k as the total number of patients in the system after the departure of a patient k, and M_k as the number of new patients arrived during the service time of the patient k, and then the discrete time process N_k constitutes a Markov chain,

$$
N_{k+1} = \begin{cases} N_k - 1 + M_{k+1} & N_k > 0, \\ M_{k+1} & N_k = 0. \end{cases}
$$

Denote

$$\widehat{N}_k = \begin{cases} N_k - 1 & N_k > 0, \\ N_k & N_k = 0. \end{cases}$$

Then

$$N_{k+1} = \widehat{N}_k + M_{k+1}.$$

Since M and $\widehat{N}$ are independent, the probability generating functions (PGFs) satisfy

$$G_N(z) = G_{\widehat{N}}(z) \cdot G_M(z).$$

To determine $G_{\widehat{N}}(z)$ and $G_M(z)$, we have

$$\begin{aligned} G_{\widehat{N}}(z) &= E[z^{\widehat{N}}] = P\{\widehat{N} = 0\} + \sum_{i=1}^{\infty} z^i P\{\widehat{N} = i\} \\ &= P\{N = 0\}\Big(1 - \frac{1}{z}\Big) + \frac{1}{z}\sum_{i=0}^{\infty} z^i P\{N = i\} \\ &= \frac{G_N(z) - (1-\rho)(1-z)}{z}, \end{aligned}$$

where $\rho = \lambda E[S]$, λ is the arrival rate, and $E[S]$ is the expectation of service time S. According to the Pollaczek-Khinchin transform equation ([136], Sec. 5.8),

$$G_M(z) = S^*((1-z)\lambda),$$

where $S^*(s)$ represents the Laplace-Stieltjes transform (LST) of S. Then

$$G_N(z) = \frac{(1-\rho)(1-z)S^*((1-z)\lambda)}{S^*((1-z)\lambda) - z}.$$

Similarly, denote $T^*(s)$ as the LST of the patient cycle time T, and then according to [136], Sec.5.8,

$$\begin{aligned} G_N(z) &= T^*((1-z)\lambda), \\ T^*(s) &= \frac{(1-\rho)sS^*(s)}{s - \lambda + \lambda S^*(s)}. \end{aligned}$$

Since the waiting time W and service time S are independent and $T = W + S$, it is natural that $T^*(s) = W^*(s)S^*(s)$, which implies that

$$W^*(s) = \frac{(1-\rho)s}{s - \lambda + \lambda S^*(s)}.$$

Taking the first and second derivatives of $W^*(s)$ at $s = 0$ and applying L'Hôspital's rule, the mean waiting time $E(W)$, the second moment $E(W^2)$, and the variance of waiting time $Var(W) = E(W^2) - E^2(W)$ can be calculated.

If server vacation is considered, denote the vacation time as V with mean $E(V)$ and LST $V^*(s)$, the PGF of N is modified as

$$\begin{aligned} G_N(z) &= \frac{(1-\rho)(1 - V^*((1-z)\lambda))S^*((1-z)\lambda)}{\lambda E(V)(S^*((1-z)\lambda) - z)} \\ &= T^*((1-z)\lambda), \end{aligned}$$

Through similar derivation, we have

$$W^*(s) = \frac{1 - V^*(s)}{sE(V)} \cdot \frac{(1-\rho)s}{s - \lambda + \lambda S^*(s)}.$$

Then, the first and second moments of the waiting time are given by

$$\begin{aligned} E(W) &= \frac{\lambda E(S^2)}{2(1-\rho)} + \frac{E(V^2)}{2E(V)}, \\ E(W^2) &= \frac{\lambda E(S^3)}{3(1-\rho)} + \frac{\lambda^2 E(S^2)^2}{2(1-\rho)^2} + \frac{\lambda E(S^2)E(V^2)}{2(1-\rho)E(V)} + \frac{E(V^3)}{3E(V)}. \end{aligned}$$

In the case of the first come first serve policy, let the total arrival rate be $\lambda = \lambda'_{ov} + \lambda'_{ev}$, and define

$$p_i = \lambda'_i/\lambda, \quad i = ev, ov.$$

Then, the first to the third moments of the service time S can be evaluated as

$$\begin{aligned}
E(S) &= \sum_{i=ev,ov} p_i \frac{1}{\mu_i} = (\frac{\lambda'_{ov}}{\mu_{ov}} + \frac{\lambda'_{ev}}{\mu_{ev}})\frac{1}{\lambda} = \frac{\rho}{\lambda}, \\
E(S^2) &= \sum_{i=ev,ov} p_i E(S_i^2) = \frac{2}{\lambda}\Big(\frac{\lambda'_{ov}\delta_{ov}}{\mu_{ov}^2} + \frac{\lambda'_{ev}\delta_{ev}}{\mu_{ev}^2}\Big) = \frac{2}{\lambda}(\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev}), \\
E(S^3) &= \sum_{i=ev,ov} p_i E(S_i^3) = \frac{\lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{\lambda}.
\end{aligned}$$

Consequently, the waiting time can be calculated as

$$E(W) = (\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev})\frac{1}{(1-\rho)} + \omega_v,$$

and the variance of waiting

$$\begin{aligned}
\mathrm{Var}(W) &= E(W^2) - E^2(W) \\
&= \frac{(\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev})^2}{(1-\rho)^2} - \omega_v^2 + \frac{(1-\rho)\mu_v E(V^3) + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho)}.
\end{aligned}$$

Finally, the average time each type of patients spend in the system and the associated variance are obtained:

$$\begin{aligned}
T_i &= E(W) + E(S_i), \\
\mathrm{Var}_i &= \mathrm{Var}(W) + \mathrm{Var}(S_i), \quad i = ev, ov.
\end{aligned}$$

Next, consider the M/G/1 queue with priorities. The formulations can be derived similarly but with more complexity. The detailed derivation can be referred to ([137], Chapters 3.3 and 3.4). For the non-preemptive policy,

$$\begin{aligned}
E(W_{ov}) &= \frac{(1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S_{ov}^2) + \lambda'_{ev}E(S_{ev}^2)}{2(1-\rho_{ov})}, \\
E(W_{ev}) &= \frac{(1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S_{ov}^2) + \lambda'_{ev}E(S_{ev}^2)}{2(1-\rho_{ov})(1-\rho)}, \\
E(W_{ov}^2) &= \frac{\lambda'_{ov}E(S_{ov}^2)(1-\rho)\frac{E(V^2)}{E(V)}}{2(1-\rho_{ov})^2} + \frac{\lambda'_{ov}E(S_{ov}^2)(\lambda'_{ov}E(S_{ov}^2) + \lambda'_{ev}E(S_{ev}^2))}{2(1-\rho_{ov})^2} \\
&\quad + \frac{(1-\rho)\frac{E(V^3)}{E(V)} + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho_{ov})}, \\
E(W_{ev}^2) &= \frac{(1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S_{ov}^2) + \lambda'_{ev}E(S_{ev}^2)}{2(1-\rho_{ov})(1-\rho)} \\
&\quad \cdot \Big(\frac{\lambda'_{ov}E(S_{ov}^2)}{(1-\rho_{ov})^2} + \frac{\lambda'_{ov}E(S_{ov}^2) + \lambda'_{ev}E(S_{ev}^2)}{(1-\rho_{ov})(1-\rho)}\Big) \\
&\quad + \frac{(1-\rho)\frac{E(V^3)}{E(V)} + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{3(1-\rho_{ov})^2(1-\rho)}.
\end{aligned}$$

The second moments are rephrased as

$$E(S_i^2) \;=\; \frac{2\omega_i}{\mu_i}, \quad i = ev, ov, v.$$

Plugging in the second moments, the expected patient length of visit and the associated variance for each type of patients can be obtained.

For the preemptive-resume policy,

$$
\begin{aligned}
E(W_{ov}) &= \frac{(1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S^2_{ov})}{2(1-\rho_{ov})}, \\
E(W_{ev}) &= \frac{(1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S^2_{ov}) + \lambda'_{ev}E(S^2_{ev})}{2(1-\rho_{ov})(1-\rho)}, \\
E(W^2_{ov}) &= \frac{\lambda'_{ov}E(S^2_{ov})((1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S^2_{ov}))}{2(1-\rho_{ov})^2} + \frac{(1-\rho)\frac{E(V^3)}{E(V)} + \lambda'_{ov}E(S^3_{ov})}{3(1-\rho_{ov})}, \\
E(W^2_{ev}) &= \frac{(1-\rho)\frac{E(V^2)}{E(V)} + \lambda'_{ov}E(S^2_{ov}) + \lambda'_{ev}E(S^2_{ev})}{2(1-\rho_{ov})(1-\rho)} \\
&\quad \cdot (\frac{\lambda'_{ov}E(S^2_{ov})}{(1-\rho_{ov})^2} + \frac{\lambda'_{ov}E(S^2_{ov}) + \lambda'_{ev}E(S^2_{ev})}{(1-\rho_{ov})(1-\rho)}) \\
&\quad + \frac{(1-\rho)\frac{E(V^3)}{E(V)} + \lambda'_{ov}E(S^3_{ov}) + \lambda'_{ev}E(S^3_{ev})}{3(1-\rho_{ov})^2(1-\rho)}.
\end{aligned}
$$

Plugging in the second moments, the expected patient length of visit and the associated variance for each type of patients can be calculated. ■

Proof of Proposition 5.1: From Theorem 5.1, take the partial derivative of T_{ov} w.r.t. β_{ov},

$$
\frac{\partial T_{ov}}{\partial \beta_{ov}} = \begin{cases}
\frac{\lambda_{ev}[\omega_{ov}+\rho_{ev}(\omega_{ev}-\omega_v)]}{\mu_{ov}(1-\rho_{ov})^2}, & \\
\qquad \text{non-preemptive policy} & \\
\frac{\lambda_{ev}[\omega_{ov}-\omega_v\rho_{ev}]}{\mu_{ov}(1-\rho_{ov})^2}, & \\
\qquad \text{preemptive-resume policy} & \\
\frac{\lambda_{ev}[\omega_{ov}(1-\rho_{ev})+\omega_{ev}\rho_{ev}]}{\mu_{ov}(1-\rho)^2}. & \\
\qquad \text{first come first serve policy} &
\end{cases}
$$

As one can see, under the first come first serve policy, $\frac{\partial T_{ov}}{\partial \beta_{ov}} > 0$ without any condition since $\rho_{ev} < 1$. Under the non-preemptive policy, $\frac{\partial T_{ov}}{\partial \beta_{ov}} > 0$ if and only if $\omega_{ov} + \rho_{ev}(\omega_{ev} -$

$\omega_v) > 0$. For the preemptive-resume policy, $\frac{\partial T_{ov}}{\partial \beta_{ov}} > 0$ if and only if $\omega_{ov} > \omega_v \rho_{ev}$. ■

Proof of Proposition 5.2: From Theorem 5.1, we obtain

$$\frac{\partial T_{ov}}{\partial \beta_{ev}} = \begin{cases} \frac{\lambda_{ws}[\beta_{ov}\mu_{ev}\omega_{ov} - (\lambda'_{ov} - \beta_{ov}\lambda_{ev} - \mu_{ov})(\omega_{ev} - \omega_v)]}{\mu_{ev}\mu_{ov}(1-\rho_{ov})^2}, \\ \qquad \text{non-preemptive policy} \\ \frac{\lambda_{ws}[\beta_{ov}\mu_{ev}\omega_{ov} + (\lambda'_{ov} - \beta_{ov}\lambda_{ev} - \mu_{ov})\omega_v]}{\mu_{ev}\mu_{ov}(1-\rho_{ov})^2}, \\ \qquad \text{preemptive-resume policy} \\ \frac{\lambda_{ws}[\omega_{ov}\rho_{ov} + \omega_{ev}(1-\rho_{ov})]}{\mu_{ev}(1-\rho)^2} + \frac{\lambda_{ws}\beta_{ov}[\omega_{ov}(1-\rho_{ev}) + \omega_{ev}\rho_{ev}]}{\mu_{ov}(1-\rho)^2}. \\ \qquad \text{first come first serve policy} \end{cases}$$

Again, $\frac{\partial T_{ov}}{\partial \beta_{ev}} > 0$ unconditionally under the first come first serve policy. After simple algebraic operations, it can be shown that under the non-preemptive policy,

$$\frac{\partial T_{ov}}{\partial \beta_{ev}} > 0 \text{ iff } \beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov} - \lambda_{ov})(\omega_v - \omega_{ev}).$$

Under the preemptive-resume policy,

$$\frac{\partial T_{ov}}{\partial \beta_{ev}} > 0 \text{ iff } \beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov} - \lambda_{ov})\omega_v.$$

■

Proof of Proposition 5.3: From Theorem 5.1, under the non-preemptive policy, we obtain:

$$\begin{aligned}\frac{\partial T_{ev}}{\partial \beta_{ov}} &= \frac{\lambda_{ev}(\omega_{ov}\rho_{ov} + \omega_{ev}\rho_{ev} + \omega_v(1-\rho))}{\mu_{ov}(1-\rho_{ov})(1-\rho)^2} + \frac{\lambda_{ev}(\omega_{ov} - \omega_v)}{\mu_{ov}(1-\rho_{ov})(1-\rho)} \\ &+ \frac{\lambda_{ev}(\omega_{ov}\rho_{ov} + \omega_{ev}\rho_{ev} + \omega_v(1-\rho))}{\mu_{ov}(1-\rho_{ov})^2(1-\rho)}, \\ \frac{\partial T_{ev}}{\partial \beta_{ev}} &= \frac{\lambda_{ws}\beta_{ov}(\omega_{ov}\rho_{ov} + \omega_{ev}\rho_{ev} + \omega_v(1-\rho))}{\mu_{ov}(1-\rho_{ov})^2(1-\rho)} \\ &+ \frac{\lambda_{ws}(\frac{\beta_{ov}\omega_{ov}}{\mu_{ov}} + \frac{\omega_{ev}}{\mu_{ev}} - \frac{\beta_{ov}\omega_v}{\mu_{ov}} - \frac{\omega_v}{\mu_{ev}})}{(1-\rho_{ov})(1-\rho)} \\ &+ \frac{\lambda_{ws}(\frac{\beta_{ov}}{\mu_{ov}} + \frac{1}{\mu_{ev}})(\omega_{ov}\rho_{ov} + \omega_{ev}\rho_{ev} + \omega_v(1-\rho))}{(1-\rho_{ov})(1-\rho)^2}.\end{aligned}$$

Simplifying the above equations, one can show that $\frac{\partial T_{ev}}{\partial \beta_{ov}} > 0$ if and only if

$$\omega_{ev}\rho_{ev}(2-\rho-\rho_{ov}) + \omega_{ov}[1-\rho+\rho_{ov}(1-\rho_{ov})] + \omega_v(1-\rho)^2 > 0,$$

which is satisfied without any condition. In addition, $\frac{\partial T_{ev}}{\partial \beta_{ev}} > 0$ if and only if

$$\begin{aligned}&\rho_{ov}(1-\rho_{ov})\omega_{ov} + (1-\rho_{ov})^2\omega_{ev} + \beta_{ov}\frac{\mu_{ev}}{\mu_{ov}}[\omega_{ev}\rho_{ev}(2-\rho-\rho_{ov}) \\ &+ \omega_{ov}[1-\rho+\rho_{ov}(1-\rho_{ov})] + \omega_v(1-\rho)^2] > 0\end{aligned}$$

which is always satisfied as well.

Under the preemptive-resume policy, we have

$$\begin{aligned}\frac{\partial T_{ev}}{\partial \beta_{ov}} &= \frac{\lambda_{ev}(\omega_{ov}\rho_{ov} + \omega_{ev}\rho_{ev} + \omega_v(1-\rho))}{\mu_{ov}(1-\rho_{ov})(1-\rho)^2} + \frac{\lambda_{ev}(\omega_{ov} - \omega_v)}{\mu_{ov}(1-\rho_{ov})(1-\rho)} \\ &+ \frac{\rho_{ev}}{\mu_{ov}(1-\rho_{ov})^2} + \frac{\lambda_{ev}(\omega_{ov}\rho_{ov} + \omega_{ev}\rho_{ev} + \omega_v(1-\rho))}{\mu_{ov}(1-\rho_{ov})^2(1-\rho)},\end{aligned}$$

$$\begin{aligned}\frac{\partial T_{ev}}{\partial \beta_{ev}} &= \frac{\lambda_{ws}\beta_{ov}(\omega_{ov}\rho_{ov}+\omega_{ev}\rho_{ev}+\omega_v(1-\rho))}{\mu_{ov}(1-\rho_{ov})^2(1-\rho)}\\ &+\frac{\lambda_{ws}(\frac{\beta_{ov}\omega_{ov}}{\mu_{ov}}+\frac{\omega_{ev}}{\mu_{ev}}-\frac{\beta_{ov}\omega_v}{\mu_{ov}}-\frac{\omega_v}{\mu_{ev}})}{(1-\rho_{ov})(1-\rho)}+\frac{\lambda_{ws}\beta_{ov}}{\mu_{ov}\mu_{ev}(1-\rho_{ov})^2}\\ &+\frac{\lambda_{ws}(\frac{\beta_{ov}}{\mu_{ov}}+\frac{1}{\mu_{ev}})(\omega_{ov}\rho_{ov}+\omega_{ev}\rho_{ev}+\omega_v(1-\rho))}{(1-\rho_{ov})(1-\rho)^2}.\end{aligned}$$

Comparing with the results under the non-preemptive policy, an additional term $\frac{\rho_{ev}}{\mu_{ov}(1-\rho_{ov})^2}$ is added to $\frac{\partial T_{ev}}{\partial \beta_{ov}}$, and an additional term $\frac{\lambda_{ws}\beta_{ov}}{\mu_{ov}\mu_{ev}(1-\rho_{ov})^2}$ is added to $\frac{\partial T_{ev}}{\partial \beta_{ev}}$. Since both of them are positive, $\frac{\partial T_{ev}}{\partial \beta_{ov}} > 0$ and $\frac{\partial T_{ev}}{\partial \beta_{ev}} > 0$ are satisfied.

Under the first come first serve policy, we have

$$\begin{aligned}\frac{\partial T_{ev}}{\partial \beta_{ov}} &= \frac{\lambda_{ev}(\omega_{ov}(1-\rho_{ev})+\omega_{ev}\rho_{ev})}{\mu_{ov}(1-\rho)^2} > 0,\\ \frac{\partial T_{ev}}{\partial \beta_{ev}} &= \frac{\lambda_{ws}(\omega_{ov}\rho_{ov}+\omega_{ev}(1-\rho_{ov}))}{\mu_{ev}(1-\rho)^2}+\frac{\lambda_{ws}\beta_{ov}(\omega_{ov}(1-\rho_{ev})+\omega_{ev}\rho_{ev})}{\mu_{ov}(1-\rho)^2} > 0.\end{aligned}$$

■

Proof of Proposition 5.4: First, consider the non-preemptive policy. Define

$$\begin{aligned}W_{ov} &= \frac{\rho_{ov}\omega_{ov}+\rho_{ev}\omega_{ev}+(1-\rho)\omega_v}{1-\rho_{ov}},\\ N_{ov} &= \frac{\rho_{ov}\omega_{ov}-\rho_{ev}\omega_{ev}-(1-\rho)\omega_v}{1-\rho_{ov}},\\ M_{ov} &= \frac{(1-\rho)\mu_v E(V^3)+\lambda'_{ov}E(S_{ov}^3)+\lambda'_{ev}E(S_{ev}^3)}{1-\rho_{ov}},\end{aligned}$$

and

$$\mathrm{Var}(W_{ov}) = W_{ov}N_{ov}+\frac{1}{3}M_{ov}.$$

The partial derivative of Var_{ov} w.r.t. β_i can be expressed as

$$\begin{aligned}\frac{\partial \mathrm{Var}_{ov}}{\partial \beta_i} &= \frac{\partial \mathrm{Var}(W_{ov})}{\partial \beta_i}\\ &= \frac{\partial W_{ov}}{\partial \beta_i}N_{ov}+W_{ov}\frac{\partial N_{ov}}{\partial \beta_i}+\frac{1}{3}\frac{\partial M_{ov}}{\partial \beta_i}, \quad i = ev, ov.\end{aligned}$$

It can be shown that

$$\begin{aligned}
\frac{\partial W_{ov}}{\partial \beta_{ov}} &> 0 \text{ iff } \omega_{ov} + \omega_{ev}\rho_{ev} - \omega_v \rho_{ev} > 0, \\
\frac{\partial N_{ov}}{\partial \beta_{ov}} &> 0 \text{ iff } \omega_{ov} - \omega_{ev}\rho_{ev} + \omega_v \rho_{ev} > 0, \\
\frac{\partial M_{ov}}{\partial \beta_{ov}} &> 0 \text{ iff } \mu_{ov} E(S_{ov}^3) + \lambda'_{ev} E(S_{ev}^3) - \rho_{ev}\mu_v E(V^3) > 0.
\end{aligned}$$

The sufficient conditions are satisfied when all of the above-mentioned inequalities are true.

Next, consider the preemptive-resume policy. Define

$$\begin{aligned}
W_{ov} &= \frac{\rho_{ov}\omega_{ov} + (1-\rho)\omega_v}{1-\rho_{ov}}, \\
N_{ov} &= \frac{\rho_{ov}\omega_{ov} - (1-\rho)\omega_v}{1-\rho_{ov}}, \\
M_{ov} &= \frac{(1-\rho)\mu_v E(V^3) + \lambda'_{ov} E(S_{ov}^3)}{1-\rho_{ov}}.
\end{aligned}$$

Similarly, one can show that $\frac{\partial W_{ov}}{\partial \beta_{ov}} > 0$ implies $\omega_{ov} > \omega_v \rho_{ev}$, $\frac{\partial M_{ov}}{\partial \beta_{ov}} > 0$ implies $\mu_{ov} E(S_{ov}^3) > \rho_{ev}\mu_v E(S_v^3)$, and $\frac{\partial N_{ov}}{\partial \beta_{ov}} > 0$ is always true, since $\omega_{ov} + \omega_v \rho_{ev} > 0$.

Finally, consider the first come first serve policy. Define

$$\begin{aligned}
W &= \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} + (1-\rho)\omega_v}{1-\rho}, \\
N &= \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} - (1-\rho)\omega_v}{1-\rho}, \\
M &= \frac{(1-\rho)\mu_v E(S_v^3) + \lambda'_{ov} E(S_{ov}^3) + \lambda'_{ev} E(S_{ev}^3)}{1-\rho}.
\end{aligned}$$

It can be shown that both $\frac{\partial W}{\partial \beta_{ov}} > 0$ and $\frac{\partial N}{\partial \beta_{ov}} > 0$ imply $\omega_{ov}(1-\rho_{ev}) + \omega_{ev}\rho_{ev} > 0$, and $\frac{\partial M}{\partial \beta_{ov}} > 0$ implies $\mu_{ov}(1-\rho_{ev})E(S_{ov}^3) + \lambda'_{ev} E(S_{ev}^3) > 0$, which are always true. ■

Proof of Proposition 5.5: First, consider the non-preemptive policy.

$$\begin{aligned}
\frac{\partial W_{ov}}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}\omega_{ov} > (\omega_v - \omega_{ev})[\beta_{ov}\rho_{ev} + \frac{\mu_{ov}}{\mu_{ev}}(1-\rho_{ov})],\\
\frac{\partial N_{ov}}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}\omega_{ov} > (\omega_{ev} - \omega_{v})[\beta_{ov}\rho_{ev} + \frac{\mu_{ov}}{\mu_{ev}}(1-\rho_{ov})],\\
\frac{\partial M_{ov}}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}[\mu_{ov}E(S_{ov}^3) + \lambda_{ev}E(S_{ev}^3) - \rho_{ev}\mu_v E(S_v^3)]\\
&\qquad + (1-\rho_{ov})\frac{\mu_{ov}}{\mu_{ev}}[\mu_{ev}E(S_{ev}^3) - \mu_v E(S_v^3)] > 0.
\end{aligned}$$

Plug-in $\rho_{ev} = \frac{\lambda_{ev}+\beta_{ev}\lambda_{ws}}{\mu_{ev}}$ and $\rho_{ov} = \frac{\lambda_{ov}+\beta_{ov}(\lambda_{ev}+\beta_{ev}\lambda_{ws})}{\mu_{ov}}$. Then, the sufficient but not necessary conditions for $\frac{\partial \text{Var}_{ov}}{\partial \beta_{ev}} > 0$ are $\beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov} - \lambda_{ov})|\omega_v - \omega_{ev}|$ and $\mu_{ev}E(S_{ev}^3) \geq \mu_v E(S_v^3)$.

Next, consider the preemptive-resume policy.

$$\begin{aligned}
\frac{\partial W_{ov}}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}(\omega_{ov} - \omega_v\rho_{ev}) - (1-\rho_{ov})\omega_v\frac{\mu_{ov}}{\mu_{ev}} > 0,\\
\frac{\partial N_{ov}}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}(\omega_{ov} + \omega_v\rho_{ev}) + (1-\rho_{ov})\omega_v\frac{\mu_{ov}}{\mu_{ev}} > 0,\\
\frac{\partial M_{ov}}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}[\mu_{ov}E(S_{ov}^3) - \rho_{ev}\mu_v E(S_v^3)] - (1-\rho_{ov})\frac{\mu_{ov}}{\mu_{ev}}\mu_v E(S_v^3) > 0.
\end{aligned}$$

The sufficient but not necessary conditions are $\frac{\partial W_{ov}}{\partial \beta_{ev}} \geq 0$ and $\frac{\partial M_{ov}}{\partial \beta_{ev}} \geq 0$, which lead to $\beta_{ov}\mu_{ev}\omega_{ov} > (\mu_{ov} - \lambda_{ov})\omega_v$ and $\beta_{ov}\mu_{ov}\mu_{ev}E(S_{ov}^3) \geq (\mu_{ov} - \lambda_{ov})\mu_v E(S_v^3)$.

Finally, consider the first come first serve policy.

$$\begin{aligned}
\frac{\partial W}{\partial \beta_{ev}} &> 0 \text{ iff } \mu_{ov}(\omega_{ov}\rho_{ov} + \omega_{ev}(1-\rho_{ov})) \\
&\qquad + \beta_{ov}\mu_{ev}[\omega_{ov}(1-\rho_{ev}) + \omega_{ev}\rho_{ev}] > 0, \\
\frac{\partial N}{\partial \beta_{ev}} &> 0 \text{ iff } \beta_{ov}(\omega_{ov}(1-\rho_{ev}) + \omega_{ev}\rho_{ev}) \\
&\qquad + \frac{\mu_{ov}}{\mu_{ev}}[\omega_{ov}\rho_{ov} + \omega_{ev}(1-\rho_{ov})] > 0, \\
\frac{\partial M}{\partial \beta_{ev}} &> 0 \text{ iff } \lambda_{ov}\mu_{ov}E(S_{ov}^3) + \mu_{ov}\mu_{ev}(1-\rho_{ov})E(S_{ev}^3) \\
&\qquad + \mu_{ev}\beta_{ov}[\mu_{ov}(1-\rho_{ev})E(S_{ov}^3) + \lambda_{ev}E(S_{ev}^3)] > 0,
\end{aligned}$$

which are always true. ■

Proof of Proposition 5.6: Consider both the non-preemptive and preemptive-resume policies. Define

$$\begin{aligned}
W_{ev} &= \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} + (1-\rho)\omega_v}{(1-\rho_{ov})(1-\rho)}, \\
N_{ev} &= \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} - (1-\rho)\omega_v}{(1-\rho_{ov})(1-\rho)} + \frac{2\rho_{ov}\omega_{ov}}{(1-\rho_{ov})^2}, \\
M_{ev} &= \frac{(1-\rho)\mu_v E(S_v^3) + \lambda'_{ov}E(S_{ov}^3) + \lambda'_{ev}E(S_{ev}^3)}{(1-\rho_{ov})^2(1-\rho)}.
\end{aligned}$$

Under the non-preemptive policy,

$$\begin{aligned}
\mathrm{Var}(W_{ev}) &= W_{ev}N_{ev} + \frac{1}{3}M_{ev}, \\
\frac{\partial \mathrm{Var}_{ev}}{\partial \beta_i} &= \frac{\partial \mathrm{Var}(W_{ev})}{\partial \beta_i} = \frac{\partial W_{ev}}{\partial \beta_i}N_{ev} + W_{ev}\frac{\partial N_{ev}}{\partial \beta_i} + \frac{1}{3}\frac{\partial M_{ev}}{\partial \beta_i}, \quad i = ev, ov.
\end{aligned}$$

In addition, under the preemptive-resume policy,

$$\begin{aligned}
A_{ev} &= \frac{2\delta_{ev} - 1}{\mu_{ev}^2(1-\rho_{ov})^2}, \\
\frac{\partial \mathrm{Var}_{ev}}{\partial \beta_i} &= \frac{\partial \mathrm{Var}(W_{ev})}{\partial \beta_i} + \frac{\partial A_{ev}}{\partial \beta_i}, \quad i = ev, ov.
\end{aligned}$$

It follows that

$$\begin{aligned}
\frac{\partial W_{ev}}{\partial \beta_{ov}} &> 0 \text{ iff } \omega_{ov}(1-\rho_{ov}^2-\rho_{ev}) \\
&\quad + \omega_{ev}\rho_{ev}(2-\rho_{ov}-\rho) + \omega_v(1-\rho)^2 > 0, \\
\frac{\partial N_{ev}}{\partial \beta_{ov}} &> 0 \text{ iff } \omega_{ov}((1-\rho_{ov}^2-\rho_{ev})(1-\rho_{ov}) + 2(1+\rho_{ov})(1-\rho)^2) \\
&\quad + \omega_{ev}\rho_{ev}(2-\rho_{ov}-\rho)(1-\rho_{ov}) - \omega_v(1-\rho)^2(1-\rho_{ov}) > 0, \\
\frac{\partial M_{ev}}{\partial \beta_{ov}} &> 0 \text{ iff } \mu_{ov}[(1+\rho_{ov})(1-\rho_{ev}) - 2\rho_{ov}^2]E(S_{ov}^3) \\
&\quad + \lambda_{ev}(3-2\rho-\rho_{ov})E(S_{ev}^3) + 2\mu_v(1-\rho)^2E(S_v^3) > 0.
\end{aligned}$$

In addition,

$$\frac{\partial A_{ev}}{\partial \beta_{ov}} = \frac{2\rho_{ev}(2\delta_{ev}-1)}{\mu_{ov}\mu_{ev}(1-\rho_{ov})^3} > 0.$$

Since all the other conditions are true except for $\frac{\partial N_{ev}}{\partial \beta_{ov}} > 0$, a sufficient but not necessary condition for $\frac{\partial \text{Var}_{ov}}{\partial \beta_{ov}} > 0$ is $\frac{\partial N_{ev}}{\partial \beta_{ov}} \geq 0$. Such a condition can be rephrased as

$$\begin{aligned}
&\omega_{ov}((1-\rho)(1-\rho_{ov}) + 2(1-\rho_{ov})(1-\rho)^2) \\
&+ 2\omega_{ev}\rho_{ev}(1-\rho)(1-\rho_{ov}) - \omega_v(1-\rho)^2(1-\rho_{ov}) \geq 0,
\end{aligned}$$

which can be further relaxed to

$$2\omega_{ov} - \omega_v \geq 0.$$

Moreover, for β_{ev},

$$\begin{aligned}
\frac{\partial W_{ev}}{\partial \beta_{ev}} > 0 \text{ iff } & \frac{\mu_{ov}}{\mu_{ev}}(1-\rho_{ov})[\rho_{ov}\omega_{ov} + (1-\rho_{ov})\omega_{ev}] \\
& + \beta_{ov}[\omega_{ov}(1-\rho_{ov}^2-\rho_{ev}) + \omega_{ev}\rho_{ev}(2-\rho_{ov}-\rho) + \omega_v(1-\rho)^2] > 0, \\
\frac{\partial N_{ev}}{\partial \beta_{ev}} > 0 \text{ iff } & \frac{\mu_{ov}}{\mu_{ev}}(1-\rho_{ov})^2[\rho_{ov}\omega_{ov} + (1-\rho_{ov})\omega_{ev}] \\
& + \beta_{ov}[\omega_{ov}((1-\rho_{ov}^2-\rho_{ev})(1-\rho_{ov}) + 2(1+\rho_{ov})(1-\rho)^2) \\
& + \omega_{ev}\rho_{ev}(2-\rho_{ov}-\rho)(1-\rho_{ov}) - \omega_v(1-\rho)^2(1-\rho_{ov})] > 0, \\
\frac{\partial M_{ev}}{\partial \beta_{ev}} > 0 \text{ iff } & \mu_{ov}[\lambda_{ov}(1-\rho_{ov}) + \beta_{ov}\mu_{ev}((1+\rho_{ov})(1-\rho_{ev}) - 2\rho_{ov}^2)]E(S_{ov}^3) \\
& + \mu_{ev}[\mu_{ov}(1-\rho_{ov})^2 + \beta_{ov}\lambda_{ev}(3-2\rho-\rho_{ov})]E(S_{ev}^3) \\
& + 2\beta_{ov}\mu_{ev}\mu_v(1-\rho)^2E(S_v^3) > 0.
\end{aligned}$$

In addition, in the case of the preemptive-resume policy,

$$\frac{\partial A_{ev}}{\partial \beta_{ev}} = \frac{2\beta_{ov}\lambda_{ws}(2\delta_{ev}-1)}{\mu_{ov}\mu_{ev}^2(1-\rho_{ov})^3} > 0.$$

Again, all the other conditions are true except for $\frac{\partial N_{ev}}{\partial \beta_{ev}} > 0$, and a sufficient but not necessary condition for $\frac{\partial \text{Var}_{ov}}{\partial \beta_{ev}} > 0$ is $\frac{\partial N_{ev}}{\partial \beta_{ev}} \geq 0$. Note that such a condition is the same as the one for $\frac{\partial N_{ev}}{\partial \beta_{ov}} \geq 0$ by adding one more positive term. Thus, the same sufficient condition applies. ■

Proof of Proposition 5.7: In the case of the first come first serve policy, since the waiting time for both the types of patients are the same, using the results from Propositions 5.4 and 5.5, $\frac{\partial \text{Var}_{ev}}{\partial \beta_{ov}} > 0$ and $\frac{\partial \text{Var}_{ev}}{\partial \beta_{ev}} > 0$ hold for any choice of parameters. ■

Proof of Proposition 5.8: First, compare $T_{\text{Non-Preemp}}$ and T_{Preemp}.

$$T_{\text{Non-Preemp}} - T_{\text{Preemp}} = \frac{\rho_{ov}\rho_{ev}(\frac{\mu_{ov}}{\mu_{ev}}\delta_{ev} - 1)}{(1-\rho_{ov})\lambda}.$$

Since $\mu_{ov} < \mu_{ev}$ and $cv_{ev} < 1$, we obtain $T_{\text{Non-Preemp}} < T_{\text{Preemp}}$. Thus, the non-preemptive service system is preferred compared with the preemptive-resume one.

Next, compare non-preemptive and first-come first-serve policies.

$$T_{\text{Non-Preemp}} - T_{\text{FCFS}} = \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} + (1-\rho)\omega_v}{(1-\rho_{ov})(1-\rho)\lambda} \cdot \lambda'_{ov}\lambda'_{ev}\Big(\frac{1}{\mu_{ov}} - \frac{1}{\mu_{ev}}\Big).$$

Since $\mu_{ov} < \mu_{ev}$, we obtain $T_{\text{Non-Preemp}} > T_{\text{FCFS}}$. Thus, the first come first serve policy yields the shortest overall patient length of visit. ■

Proof of Proposition 5.9: For the variance of the non-preemptive and the preemptive-resume policies,

$$\begin{aligned}\text{Var}_{\text{Non-Preemp}} - \text{Var}_{\text{Preemp}} &= p_{ov}\Big(\frac{\lambda_{ev}E(S_{ev}^3)}{3(1-\rho_{ov})} - \frac{\rho_{ev}\omega_{ev}(\rho_{ev}\omega_{ev} + 2(1-\rho)\omega_v)}{(1-\rho_{ov})^2}\Big)\\ &+ p_{ev}\Big(1 - \frac{1}{(1-\rho_{ov})^2}\Big)\Big(\frac{2\delta_{ev}-1}{\mu_{ev}^2}\Big).\end{aligned}$$

It follows that $\text{Var}_{\text{Non-Preemp}} - \text{Var}_{\text{Preemp}} < 0$ if and only if

$$\begin{aligned}&3\rho_{ov} - 6 + \mu_{ev}[12 - 6\rho_{ov} + 6(1-\rho)\mu_{ov}\omega_v]\omega_{ev}\\ &+ 3\lambda'_{ev}\mu_{ov}\omega_{ev}^2 - (1-\rho_{ov})\mu_{ov}\mu_{ev}^2E(S_{ev}^3) > 0.\end{aligned}$$

Here, we consider the third distribution moment of the e-visit service, and the condition becomes

$$E(S_{ev}^3) < \frac{1}{(1-\rho_{ov})\mu_{ov}\mu_{ev}^2}(3\rho_{ov} - 6 + \mu_{ev}[12 - 6\rho_{ov} + 6(1-\rho)\mu_{ov}\omega_v]\omega_{ev} + 3\lambda'_{ev}\mu_{ov}\omega_{ev}^2).$$

■

Proof of Corollary 5.5: When all the service and vacation time distributions are exponential and $\mu_{ov} < \mu_{ev}$,

$$\begin{aligned}
&\text{Var}^{\text{exp}}_{\text{Non-Preemp}} - \text{Var}^{\text{exp}}_{\text{Preemp}} < 0 \text{ iff} \\
&2(1-\rho)\lambda'_{ov}\mu_{ev} + [(2-\rho_{ov})\rho_{ov}\mu_{ev} - \lambda'_{ov}(2-\rho_{ov}-\rho)]\mu_v > 0.
\end{aligned}$$

Let $\lambda'_{ov} = a\lambda'_{ev}$, $\mu_{ov} = b\mu_{ev}$, and $b < 1$,

$$\begin{gathered}
(2-\rho_{ov})\rho_{ov}\mu_{ev} - \lambda'_{ov}(1-\rho_{ov}+1-\rho) > 0 \text{ iff} \\
\Big(-2a + \frac{2a}{b}\Big) + \Big(a - \frac{a^2}{b^2} + \frac{2a^2}{b}\Big)\rho_{ev} > 0.
\end{gathered}$$

If $(a - \frac{a^2}{b^2} + \frac{2a^2}{b}) \geq 0$, the above inequality is satisfied. If not, the above inequality can be written as

$$\Big(2a - \frac{2a}{b}\Big) + \Big(-a + \frac{a^2}{b^2} - \frac{2a^2}{b}\Big)\rho_{ev} < 0.$$

Since $\rho < 1$, which implies that $1-(1+\frac{a}{b})\rho_{ev} > 0$, it leads to $\rho_{ev} < \frac{b}{a+b}$. Then we obtain

$$\begin{aligned}
&\Big(2a - \frac{2a}{b}\Big) + \Big(-a + \frac{a^2}{b^2} - \frac{2a^2}{b}\Big)\rho_{ev} \\
&< 2\Big(a - \frac{a}{b}\Big) + \Big(-a + \frac{a^2}{b^2} - \frac{2a^2}{b}\Big)\frac{b}{a+b} \\
&= -\frac{a[a+(2-b)b]}{b(a+b)} < 0.
\end{aligned}$$

Next, compare Var_{FCFS} and $\text{Var}_{\text{Non-Preemp}}$. Define

$$\begin{aligned}
&M = (1-\rho)\mu_v E(V^3) + \lambda'_{ov}E(S^3_{ov}) + \lambda'_{ev}E(S^3_{ev}), \\
&N = \rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} + (1-\rho)\omega_v.
\end{aligned}$$

Then the difference in variance between the two policies:

$$\begin{aligned}
&\mathrm{Var}^{\mathrm{exp}}_{\text{Non-Preemp}} - \mathrm{Var}^{\mathrm{exp}}_{\text{FCFS}} \\
&= \Big(\frac{p_{ov}}{1-\rho_{ov}} + \frac{p_{ev}}{(1-\rho_{ov})^2(1-\rho)} - \frac{1}{1-\rho}\Big)\frac{M}{3} + \Big[p_{ov}\frac{\rho_{ov}\omega_{ov} - \rho_{ev}\omega_{ev} - (1-\rho)\omega_v}{(1-\rho_{ov})^2} \\
&+ p_{ev}\Big(\frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} - (1-\rho)\omega_v}{(1-\rho_{ov})^2(1-\rho)^2} + \frac{2\rho_{ov}\omega_{ov}}{(1-\rho_{ov})^3(1-\rho)}\Big) \\
&- \frac{\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev} - (1-\rho)\omega_v}{(1-\rho)^2}\Big]N.
\end{aligned}$$

■

Proofs of Chapter 6

Proof of Proposition 6.1: Recall (6.4),

$$\rho = \rho_{ov} + \rho_{ev} = (1 - \alpha(1 - \gamma - \beta_{ov}))\rho_t.$$

Since $0 < \alpha < 1$, then, $\rho < \rho_t$ if and only if $1 - \gamma - \beta_{ov} > 0$. ■

Proof of Proposition 6.2: Compare the change in the average cycle time $T_{ov} - T_t$:

$$T_{ov} - T_t = \frac{\alpha\lambda\left(\gamma\delta_{ev}\left(-\lambda + \gamma\mu_{ev}\right) + \delta_{ov}\left(\lambda - (1-\beta_{ov})\mu_{ev}\right)\right)}{\mu_{ev}\left(\lambda - \gamma\mu\right)\left((1 - \alpha(1-\gamma-\beta_{ov}))\lambda - \gamma\mu_{ev}\right)}. \tag{A.3}$$

According to the physician utilization constraints, the denominator of (A.3) is larger than zero, so a closer look is taken at the numerator:

$$\begin{aligned}
&\gamma\delta_{ev}(-\lambda+\gamma\mu_{ev})+\delta_{ov}(\lambda-(1-\beta_{ov})\mu_{ev})\\
&\leq \gamma\delta_{ov}(-\lambda+\gamma\mu_{ev})+\delta_{ov}(\lambda-(1-\beta_{ov})\mu_{ev})\\
&=((1-\gamma)\lambda-(1-\gamma^2-\beta_{ov})\mu_{ev})\delta_{ov}\\
&<((1-\gamma)\gamma\mu_{ev}-(1-\gamma^2-\beta_{ov})\mu_{ev})\delta_{ov}\\
&=-(1-\gamma-\beta_{ov})\mu_{ev}\delta_{ov}.
\end{aligned}$$

The above inequalities indicate that a sufficient but not necessary condition for $T_{ov}-T_t<0$ is $1-\gamma-\beta_{ov}\geq 0$. ■

Proof of Proposition 6.3: Under the same total external arrival rate λ, if $1-\gamma-\beta_{ov}=0$,

$$\begin{aligned}
\Phi_{\beta_{ov}=1-\gamma} &= -\alpha\lambda(1-\gamma)\omega_v\\
&+ \alpha\lambda^2\frac{(-1+2\gamma+(1-\gamma)\gamma\alpha)\delta_{ov}-\gamma^2(1+(1-\gamma)\alpha)\delta_{ev}}{\gamma\mu_{ev}(-\lambda+\gamma\mu_{ev})}.
\end{aligned}$$

Therefore, a necessary and sufficient condition for $\Phi_{\beta_{ov}=1-\gamma}>0$ is

$$\omega_v<\lambda\frac{(-1+2\gamma+(1-\gamma)\gamma\alpha)\delta_{ov}-\gamma^2(1+(1-\gamma)\alpha)\delta_{ev}}{\gamma(1-\gamma)\mu_{ev}(-\lambda+\gamma\mu_{ev})}. \tag{A.4}$$

Then, according to the monotonicity of Φ with respect to β_{ov}, if the condition (A.4) is satisfied, then when $1-\gamma-\beta_{ov}>0$, $\Phi>0$. ■

Proof of Proposition 6.4: When the physician's other nondirect care work is not considered, the system can be modeled as a single server queue. Specifically, when e-visits are not offered,

$$T_t=\frac{\rho_t\omega_{ov}}{1-\rho_t}+\frac{1}{\mu_{ov}}. \tag{A.5}$$

When part of patients use e-visits,

$$T_{ev} = (\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev})\frac{1}{(1-\rho)} + \frac{1}{\mu_{ev}}, \tag{A.6}$$

$$T_{ov} = (\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev})\frac{1}{(1-\rho)} + \frac{1}{\mu_{ov}}. \tag{A.7}$$

Using (A.5)-(A.7) to calculate Φ_{nv}, it can be shown that $\Phi = \Phi_{nv} - \alpha\beta_{ov}\omega_v$. ■

Proof of Proposition 6.5: Take the partial derivatives of Φ with respect to e-visit service rate μ_{ev} and the variation factor δ_{ev}:

$$\begin{aligned}\frac{\partial\Phi}{\partial\mu_{ev}} &= \frac{\rho_{ev}}{\mu_{ev}} + (\frac{2\rho_{ev}\omega_{ev}}{(1-\rho)\mu_{ev}} + \frac{(\rho_{ev}\omega_{ev} + \rho_{ov}\omega_{ov})\rho_{ev}}{(1-\rho)^2\mu_{ev}})(1+\alpha\beta_{ov})\lambda > 0,\\ \frac{\partial\Phi}{\partial\delta_{ev}} &= -\frac{\rho_{ev}(1+\alpha\beta_{ov})\lambda}{(1-\rho)\mu_{ev}} < 0.\end{aligned}$$

Therefore, Φ is monotonically increasing with respect to e-visit service rate μ_{ev}, and is monotonically decreasing with respect to e-visit variation factor δ_{ev}. ■

Proof of Proposition 6.6: Take the partial derivative of Φ with respect to β_{ov},

$$\begin{aligned}\frac{\partial\Phi}{\partial\beta_{ov}} &= -\frac{\alpha\lambda}{\mu_{ov}} - \alpha\lambda\omega_v - \frac{\alpha}{\mu_{ov}}\frac{(1+\alpha\beta_{ov})\lambda^2}{1-\rho}\omega_{ov}\\ &- (\frac{\alpha\lambda}{1-\rho} + \frac{\alpha}{\mu_{ov}}\frac{(1+\alpha\beta_{ov})\lambda^2}{(1-\rho)^2})(\rho_{ov}\omega_{ov} + \rho_{ev}\omega_{ev}) < 0.\end{aligned}$$

It can be concluded that Φ is monotonically decreasing with β_{ov}. ■

Proof of Corollary 6.2: Suppose $\beta = 1-\gamma$ and plug-in $\delta_{ov} = \delta_{ev}$,

$$\Phi_{\beta_{ov}=1-\gamma} = -\alpha\lambda(1-\gamma)\omega_v - \frac{\alpha\lambda^2(1-\gamma)^2(1-\gamma\alpha)\delta_{ov}}{\gamma\mu_{ev}(-\lambda+\gamma\mu_{ev})} < 0.$$

Based on the monotonicity property of Φ with respect to β_{ov}, when $\beta_{ov} \geq 1-\gamma$, $\Phi_{\delta_{ov}=\delta_{ev}} < 0$. ■

Proof of Proposition 6.7: According to the physician utilizations defined in (6.15) and (6.16),

$$\rho - \rho_t = \frac{(1-\alpha(1-\gamma-\beta_{ov}))\lambda - \lambda_t}{\gamma\mu_{ev}} \leq 0, \text{ iff.}$$
$$\lambda \leq \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}.$$

Thus, the maximum arrival rate λ^* the physician can accommodate is

$$\lambda^* = \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}.$$

■

Proof of Proposition 6.8: Plug-in $\lambda = \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}$ to (6.6),

$$\begin{aligned} T_{ov} - T_t &= -\frac{\alpha\left(\delta_{ov} - \gamma\delta_{ev}\right)\lambda_t}{\mu_{ev}(1-\alpha(1-\gamma-\beta_{ov}))\left(\gamma\mu_{ev} - \lambda_t\right)} < 0, \\ T_{ev} - T_t &< T_{ov} - T_t < 0. \end{aligned}$$

■

Proof of Proposition 6.9: When $\lambda = \frac{\lambda_t}{1-\alpha(1-\gamma-\beta_{ov})}$ and $\delta_{ov} = \delta_{ev}$,

$$\Phi_{\lambda=\lambda^*} = -\frac{\alpha(1-\gamma)^2\delta_{ov}(1-\alpha(1-\beta_{ov}))\lambda_t^2}{\gamma\mu_{ev}(1-\alpha(1-\gamma-\beta_{ov}))^2\left(\gamma\mu_{ev}-\lambda_t\right)} - \frac{\alpha(1-\gamma)\lambda_t\omega_v}{1-\alpha(1-\gamma-\beta_{ov})} < 0.$$

■

Proof of Corollary 6.3: In (6.22), the denominator is a quadratic function of α. To find the optimal α^* to maximize λ^* is equal to maximize $\alpha(1-\gamma-k_1\alpha-k_2)$, which yields

$$\alpha^* = \frac{1-\gamma-k_2}{2k_1}.$$

■

Derivations in Chapter 7

Consider the state space defined by assumptions (i)-(vi) in Section 7.1.2. Let $S_k = \{s_1^k$, s_2^k, s_3^k, s_4^k, s_5^k, $s_6^k\}$, where s_i^k represents the number of patients in stage i and state k, $i = 1, \ldots, 6$. $s_2^k = m$ indicates that there are m patients in process 2 (rooming) in state k. However, not every state is feasible due to the constraints. To figure out how many states are feasible, the following constraints are considered:

- $s_1^k \leq Q$, queue length constraint,
- $s_2^k + s_3^k + s_4^k + s_5^k + s_6^k \leq M$, room space constraint,
- $s_2^k + s_6^k \leq r_1$, clinical staff resource constraint,
- $s_4^k \leq r_2$, clinician resource constraint.

In addition, for any feasible state, we have

- $s_3^k > 0$ only when $s_4^k = r_2$ (the clinician is busy),
- $s_5^k > 0$ only when $s_2^k + s_6^k = r_1$ (the clinical staff is busy),
- $s_1^k > 0$ only when $\sum_{i=2}^{6} s_i^k = M$ (all the rooms are occupied), or $s_2^k + s_6^k = r_1$ (the clinical staff is occupied).

Therefore, the number of feasible states, K, is reduced. It can be described by a function of the maximum queue length Q and the number of exam rooms M, and can be obtained using numerical search. The current GI clinic can be characterized by 79 feasible states. Denote a feasible state as S_k, $k = 1, \ldots, K$. The steady state probability for a feasible state S_k is then defined as

$$P_k = P(s_1^k, s_2^k, s_3^k, s_4^k, s_5^k, s_6^k), \quad k = 1, 2, \ldots, K.$$

For a feasible state S_j, there may exist a transition from another feasible state S_k to S_j, triggered by one of the following events: (1) a patient arrives; (2) patient rooming finishes; (3) clinician examination finishes; and (4) a patient checks out. Note that such events cannot occur simultaneously. The transition rates of these events are outlined below:

- For patient arrival, the following scenarios exist:
  - If $s_1^k = Q$ (i.e., the waiting room is full, although unlikely), the patient is lost due to the space limit and no transition occurs.
  - Otherwise, the transition rate is λ.
- The clinical staff finishes rooming, and the transition rate is c_1.
- The clinician finishes examination and diagnosis, and the transition rate is c_2.
- The patient checks out, and the transition rate is c_3.

Let $I_{\{X\}}$ be the indicator of whether event X occurs or not, i.e.,

$$I_{\{X\}} = \begin{cases} 1, \text{ if } X \text{ is true,} \\ 0, \text{ if } X \text{ is false.} \end{cases} \tag{A.8}$$

Then, for a feasible state S_k, the rate goes out of S_k can be written as

$$\begin{aligned} \mu_{out}^k &= \{\lambda \cdot I_{\{0 \leq s_1^k < Q\}} + c_1 \cdot I_{\{s_2^k > 0\}} + c_2 \cdot I_{\{s_4^k > 0\}} \\ &\quad + c_3 \cdot I_{\{s_6^k > 0\}}\} \cdot P(s_1^k, s_2^k, s_3^k, s_4^k, s_5^k, s_6^k), \end{aligned} \tag{A.9}$$

which represents that any one of the four events described above could trigger the leave from state S_k.

Similarly, the rate goes into S_k can be written as:

$$\begin{aligned}
\mu_{in}^k &= c_1 \cdot [P(s_1^k, s_2^k+1, s_3^k, s_4^k-1, s_5^k+1, s_6^k-1) + P(s_1^k, s_2^k+1, s_3^k, s_4^k-1, s_5^k, s_6^k) \\
&\quad +P(s_1^k, s_2^k+1, s_3^k-1, s_4^k, s_5^k+1, s_6^k-1) + P(s_1^k, s_2^k+1, s_3^k-1, s_4^k, s_5^k, s_6^k) \\
&\quad +P(s_1^k+1, s_2^k, s_3^k, s_4^k-1, s_5^k, s_6^k) + P(s_1^k+1, s_2^k, s_3^k-1, s_4^k, s_5^k, s_6^k)] \\
&\quad +c_2 \cdot [P(s_1^k, s_2^k, s_3^k+1, s_4^k, s_5^k, s_6^k-1) + P(s_1^k, s_2^k, s_3^k, s_4^k+1, s_5^k, s_6^k-1) \\
&\quad +P(s_1^k, s_2^k, s_3^k+1, s_4^k, s_5^k-1, s_6^k) + P(s_1^k, s_2^k, s_3^k, s_4^k+1, s_5^k-1, s_6^k)] \\
&\quad +c_3 \cdot [P(s_1^k, s_2^k, s_3^k, s_4^k, s_5^k+1, s_6^k) + P(s_1^k+1, s_2^k-1, s_3^k, s_4^k, s_5^k, s_6^k+1) \\
&\quad +P(s_1^k, s_2^k, s_3^k, s_4^k, s_5^k, s_6^k+1)] + \lambda \cdot [P(s_1^k-1, s_2^k, s_3^k, s_4^k, s_5^k, s_6^k) \\
&\quad +P(s_1^k, s_2^k-1, s_3^k, s_4^k, s_5^k, s_6^k)]. \qquad \text{(A.10)}
\end{aligned}$$

In (A.10), the first bracket includes the possible events that happen after a patient finishes rooming, and the corresponding transition rate is c_1. Note that for a certain state S_k, the transition is triggered by one event at one time so that only one of the corresponding probabilities in the bracket can be nonzero. The next three brackets denote the probabilities of events that are triggered by the clinician finishing examination, the patient check-out, and a new patient arrival, with transition rates c_2, c_3 and λ, respectively. Again, only one event can trigger the transition in each bracket. Finally, the balance equations can be written as

$$\mu_{in}^k = \mu_{out}^k, \quad k = 1, 2, \ldots, K. \qquad \text{(A.11)}$$

The transitions and the corresponding transition rates from state S_k to state S_j are elaborated below:

1. For patient arrival, the following scenarios exist:

   - If $s_1^k = Q$ (i.e., the waiting room is full, although unlikely), the patient is lost due to the space limit and no transition occurs.

- If $s_1^k < Q$ (there's available space in the waiting room) and $\sum_{i=2}^6 s_i^k = M$ (i.e., all rooms are occupied by the patients), then the patient has to wait, we have

$$s_1^j = s_1^k + 1, \quad s_i^j = s_i^k, \quad i = 2, \ldots, 6.$$

- If $\sum_{i=2}^6 s_i^k < M$ (i.e, not all rooms are occupied) and $s_2^k + s_6^k < n_1$ (the clinical staff is available), then the patient does not need to wait and will be roomed immediately,

$$s_2^j = s_2^k + 1, \quad s_i^j = s_i^k, \quad i = 1, 3, \ldots, 6.$$

- However, if $\sum_{i=2}^6 s_i^k < M$ and $s_2^k + s_6^k = n_1$ (i.e., the clinical staff is busy), then the patient still needs to wait,

$$s_1^j = s_1^k + 1, \quad s_i^j = s_i^k, \quad i = 2, \ldots, 6.$$

In all three latter cases, the transition rate is λ.

2. When the clinical staff finishes rooming, the following scenarios need to be considered:

   - If $\sum_{i=2}^6 s_i^k = M$ (all rooms are occupied), then

$$s_1^j = s_1^k, \quad s_2^j = s_2^k - 1.$$

     - If $s_4^k < n_2$ (the clinician is available),

$$s_3^j = s_3^k, \quad s_4^j = s_4^k + 1.$$

       Furthermore, since the resource (clinical staff) is released,

    - ∗ if $s_5^k > 0$ (there are patients waiting for check-out),

      $$s_5^j = s_5^k - 1, \quad s_6^j = s_6^k + 1,$$

    - ∗ otherwise,

      $$s_5^j = s_5^k, \quad s_6^j = s_6^k.$$

  - If $s_4^k = n_2$ (the clinician is busy),

    $$s_3^j = s_3^k + 1, \quad s_4^j = s_4^k.$$

    The changes in processes 5 and 6 are the same as in the previous scenario ($s_5^k > 0$).

- If $\sum_{i=2}^{6} s_i^k < M$ (not all rooms are occupied),

  - if $s_4^k < n_2$ (the clinician is available),

    $$s_3^j = s_3^k, \quad s_4^j = s_4^k + 1.$$

    Moreover,

    - ∗ if $s_5^k > 0$ (there are patients waiting for check-out),

      $$\begin{aligned} s_1^j &= s_1^k, & s_2^j &= s_2^k - 1, \\ s_5^j &= s_5^k - 1, & s_6^j &= s_6^k + 1, \end{aligned}$$

    - ∗ if $s_5^k = 0$ (there is no patient waiting for check-out) and $s_1^k > 0$ (there are patients waiting for rooming),

      $$\begin{aligned} s_1^j &= s_1^k - 1, & s_2^j &= s_2^k, \\ s_5^j &= s_5^k, & s_6^j &= s_6^k, \end{aligned}$$

- $*$ if no patient is waiting for rooming or discharge,

$$s_2^j = s_2^k - 1, \quad s_i^j = s_i^k, \quad i = 1, 5, 6.$$

- If $s_4^k = n_2$ (the clinician is busy),

$$s_3^j = s_3^k + 1, \quad s_4^j = s_4^k.$$

The changes in processes 1, 2, 5, and 6 are the same as in the previous case ($s_4^k < n_2$).

In all above scenarios, the transition rate is c_1.

3. If the clinician finishes examining, $s_1^j = s_1^k$, $s_2^j = s_2^k$, then

- if $s_2^k + s_6^k < n_1$ (the clinical staff is available),

$$s_5^j = s_5^k, \quad s_6^j = s_6^k + 1,$$

In addition,

- if there are patients waiting for the clinician,

$$s_3^j = s_3^k - 1, \quad s_4^j = s_4^k,$$

- if there is no patient waiting for the clinician,

$$s_3^j = s_3^k, \quad s_4^j = s_4^k - 1.$$

- If $s_2^k + s_6^k = n_1$ (the clinical staff is busy),

$$s_5^j = s_5^k + 1, \quad s_6^j = s_6^k.$$

The changes in processes 3 and 4 are the same as in the previous case ($s_2^k + s_6^k < n_1$).

In all these cases, the transition rate is c_2.

4. If any patient checks out, then

$$s_3^j = s_3^k, \quad s_4^j = s_4^k.$$

In addition, the following scenarios exist:

- If there are patients waiting for check-out,

$$\begin{aligned} s_1^j &= s_1^k, & s_2^j &= s_2^k, \\ s_5^j &= s_5^k - 1, & s_6^j &= s_6^k. \end{aligned}$$

- If there is no patient waiting for check-out, but
  - there are patients waiting for rooming,

$$\begin{aligned} s_1^j &= s_1^k - 1, & s_2^j &= s_2^k + 1, \\ s_5^j &= s_5^k, & s_6^j &= s_6^k - 1, \end{aligned}$$

  - otherwise,

$$\begin{aligned} s_1^j &= s_1^k, & s_2^j &= s_2^k, \\ s_5^j &= s_5^k, & s_6^j &= s_6^k - 1. \end{aligned}$$

Again, in these scenarios, the transition rate is c_3.

■

Bibliography

[1] National Association of Community Health Centers. Access is the answer: Community health centers, primary care & the future of american health care. 2014.

[2] Sage M Timberline. It takes a village: The importance of a comprehensive definition of primary healthcare access for just and effective policy. 2015.

[3] Clese E Erikson, Sana Danish, Karen C Jones, Shana F Sandberg, and Adam C Carle. The role of medical school culture in primary care career choice. *Academic Medicine*, 88(12):1919–1926, 2013.

[4] Naveen Gidwani, Louis Fernandez, and David Schlossman. Connecting with patients online: E-visits. *Consulting report prepared for the US Department of Family and Community Medicine Academic Health Center*.

[5] Adele Marshall, Christos Vasilakis, and Elia El-Darzi. Length of stay-based patient flow models: recent developments and future directions. *Health Care Management Science*, 8(3):213–220, 2005.

[6] Randolph Hall. *Patient flow: reducing delay in healthcare delivery*, volume 206. Springer Science & Business Media, 2013.

[7] Jong Jun, Sheldon H Jacobson, and James R Swisher. Application of discrete-event simulation in health care clinics: A survey. *Journal of the Operational Research Society*, 50(2):109–123, 1999.

[8] Sheldon H Jacobson, Shane N Hall, and James R Swisher. Discrete-event simulation of health care systems. In *Patient flow: Reducing delay in healthcare delivery*, pages 211–252. Springer, 2006.

[9] Sally C Brailsford. Advances and challenges in healthcare simulation modeling: tutorial. In *Proceedings of the 39th conference on Winter simulation*, pages 1436–1448. Winter Simulation Conference, 2007.

[10] E El-Darzi, C Vasilakis, T Chaussalet, and PH Millard. A simulation modelling approach to evaluating length of stay, occupancy, emptiness and bed blocking in a hospital geriatric department. *Health Care Management Science*, 1(2):143–149, 1998.

[11] Simon Samaha, Wendy S Armel, and Darrell W Starks. Emergency departments i: the use of simulation to reduce the length of stay in an emergency department. In *Proceedings of the 35th conference on Winter simulation*, pages 1907–1911. Winter Simulation Conference, 2003.

[12] Alexander Komashie and Ali Mousavi. Modeling emergency departments using discrete event simulation techniques. In *Proceedings of the 37th conference on Winter simulation*, pages 2681–2685. Winter Simulation Conference, 2005.

[13] Christine Duguay and Fatah Chetouane. Modeling and improving emergency department systems using discrete event simulation. *Simulation*, 83(4):311–320, 2007.

[14] Nathan R Hoot, Larry J LeBlanc, Ian Jones, Scott R Levin, Chuan Zhou, Cynthia S Gadd, and Dominik Aronsky. Forecasting emergency department crowding: a discrete event simulation. *Annals of Emergency Medicine*, 52(2):116–125, 2008.

[15] Alexander Kolker. Process modeling of emergency department patient flow: Effect of patient length of stay on ed diversion. *Journal of Medical Systems*, 32(5):389–401, 2008.

[16] Mariagrazia Dotoli, Maria Pia Fanti, Agostino M Mangini, and Walter Ukovich. A continuous petri net model for the management and design of emergency cardiology

departments. In *Analysis and Design of Hybrid Systems*, volume 3, pages 50–55, 2009.

[17] Stuart Brenner, Zhen Zeng, Yang Liu, Junwen Wang, Jingshan Li, and Patricia K Howard. Modeling and analysis of the emergency department at university of kentucky chandler hospital using simulations. *Journal of Emergency Nursing*, 36(4):303–310, 2010.

[18] 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.

[19] Zhen Zeng, Xiaoji Ma, Yao Hu, Jingshan Li, and Deborah Bryant. A simulation study to improve quality of care in the emergency department of a community hospital. *Journal of Emergency Nursing*, 38(4):322–328, 2012.

[20] Samuel Fomundam and Jeffrey W Herrmann. A survey of queuing theory applications in healthcare. *The Institute for Systems Research Technical Report 2007-24*, 2007.

[21] Linda V Green, Joao Soares, James F Giglio, and Robert A Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[22] Susanna WM Au-Yeung, Peter G Harrison, and William J Knottenbelt. Approximate queueing network analysis of patient treatment times. In *Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools*, page 45. Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 2007.

[23] L Mayhew and D Smith. Using queuing theory to analyse the governments 4-h completion time target in accident and emergency departments. *Health Care Management Science*, 11(1):11–21, 2008.

[24] 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.

[25] Steven C Adamson and John W Bachman. Pilot study of providing online care in a primary care setting. In *Mayo Clinic Proceedings*, volume 85, pages 704–710. Elsevier, 2010.

[26] David Baer. Patient-physician e-mail communication: the kaiser permanente experience. *Journal of Oncology Practice*, 7(4):230–233, 2011.

[27] Jennifer Prestigiacomo. Making the evisit work. *Healthcare Informatics Online, http://www.healthcare-informatics.com/article/making-evisit-work*, 2012.

[28] Pamela Whitten, Lorraine Buis, and Brad Love. Physician-patient e-visit programs. *Disease Management & Health Outcomes*, 15(4):207–214, 2007.

[29] Denise Scott. Putting the e-visit to test: From concept to clinic. *Academy Health*, 2010.

[30] Yi Yvonne Zhou, Michael H Kanter, Jian J Wang, and Terhilda Garrido. Improved quality at kaiser permanente through e-mail between physicians and patients. *Health Affairs*, 29(7):1370–1375, 2010.

[31] Ateev Mehrotra, Hangsheng Liu, John L Adams, Margaret C Wang, Judith R Lave, N Marcus Thygeson, Leif I Solberg, and Elizabeth A McGlynn. Comparing costs and quality of care at retail clinics with that of other medical settings for 3 common illnesses. *Annals of Internal Medicine*, 151(5):321–328, 2009.

[32] Alice J Watson, Hagit Bergman, Christy M Williams, and Joseph C Kvedar. A randomized trial to evaluate the efficacy of online follow-up visits in the management of acne. *Archives of Dermatology*, 146(4):406–411, 2010.

[33] Ateev Mehrotra, Suzanne Paone, G Daniel Martich, Steven M Albert, and Grant J Shevchik. A comparison of care at e-visits and physician office visits for sinusitis and urinary tract infection. *JAMA Internal Medicine*, 173(1):72–74, 2013.

[34] Laurence Baker, Jeffrey Rideout, Paul Gertler, and Kristiana Raube. Effect of an internet-based system for doctor-patient communication on health care spending. *Journal of the American Medical Informatics Association*, 12(5):530–536, 2005.

[35] Thomas K Houston, Daniel Z Sands, Mollie W Jenckes, and Daniel E Ford. Experiences of patients who were early adopters of electronic communication with their physician: satisfaction, benefits, and concerns. *The American Journal of Managed Care*, 10(9):601–608, 2004.

[36] Barak Gaster, Christopher L Knight, Dawn E Witt, John VL Sheffield, Nassim P Assefi, and Dedra Buchwald. Physicians' use of and attitudes toward electronic mail for patient communication. *Journal of General Internal Medicine*, 18(5):385–389, 2003.

[37] Devon M Herrick. Telemedicine provides benefits, but security and privacy risks abound. *Health Care News. http://www.heartland.org/Article.cfm*, 2006.

[38] Presidents Council of Advisors on Science and Technology. Report to the president: Better health care and lower costs: Accelerating improvement through systems engineering. 2014.

[39] Rupa Sheth Valdez, Edmond Ramly, and Patricia Flatley Brennan. Final report: Industrial and systems engineering and health care: Critical areas of research. *Agency for Healthcare Research and Quality Publication No. 10-0079-EF*, 2010.

[40] Rob Ryan. Primary care redesign. *Clinical Contributions*, 21(2):33, 1997.

[41] American College of Physicians. The impending collapse of primary care medicine and its implications for the state of the nations health care. 2006.

[42] University of Wisconsin Health. Redesigning primary care: Partnering with patients to improve health. *UW Health Technical Report*, 2013.

[43] John W Beasley, Pascale Carayon, and Mindy A Simith. Improving the quality and efficiency of primary care through industrial and systems engineering–a white paper. 2013.

[44] Karen Davis, Stephen C Schoenbaum, and Anne-Marie Audet. A 2020 vision of patient-centered primary care. *Journal of General Internal Medicine*, 20(10):953–957, 2005.

[45] Thomas Bodenheimer. Primary carewill it survive? *New England Journal of Medicine*, 355(9):861–864, 2006.

[46] Thomas Bodenheimer and Hoangmai H Pham. Primary care: current problems and proposed solutions. *Health Affairs*, 29(5):799–805, 2010.

[47] 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.

[48] Paul Bower, S Campbell, Chris Bojke, and Bonnie Sibbald. Team structure, team climate and the quality of care in primary care: an observational study. *Quality and Safety in Health Care*, 12(4):273–279, 2003.

[49] Kevin Grumbach and Thomas Bodenheimer. Can health care teams improve primary care practice? *The Journal of the American Medical Association*, 291(10):1246–1251, 2004.

[50] Geoffrey K Mitchell, Jennifer J Tieman, and Tania M Shelby-James. Multidisciplinary care planning and teamwork in primary care. *Medical Journal of Australia*, 188(8):s61–s64, 2008.

[51] Jürgen Unützer, Wayne Katon, Christopher M Callahan, John W Williams Jr, Enid Hunkeler, Linda Harpole, Marc Hoffing, Richard D Della Penna, Polly Hitchcock Noël, Elizabeth HB Lin, et al. Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. *The Journal of the American Medical Association*, 288(22):2836–2845, 2002.

[52] Michael Leonard, Suzanne Graham, and Doug Bonacum. The human factor: the critical importance of effective teamwork and communication in providing safe care. *Quality and Safety in Health Care*, 13(1):i85–i90, 2004.

[53] Adrienne Shaw, Simon de Lusignan, and George Rowlands. Do primary care professionals work as a team: A qualitative study. *Journal of Interprofessional Care*, 19(4):396–405, 2005.

[54] Simon de Lusignan and Chris van Weel. The use of routinely collected computer data for research in primary care: opportunities and challenges. *Family Practice*, 23(2):253–263, 2006.

[55] Dave A Ludwick and John Doucette. Adopting electronic medical records in primary care: lessons learned from health information systems implementation experience in seven countries. *International Journal of Medical Informatics*, 78(1):22–31, 2009.

[56] Krish Thiru, Alan Hassey, and Frank Sullivan. Systematic review of scope and quality of electronic patient record data in primary care. *British Medical Journal*, 326(7398):1070, 2003.

[57] Richard Hillestad, James Bigelow, Anthony Bower, Federico Girosi, Robin Meili, Richard Scoville, and Roger Taylor. Can electronic medical record systems transform health care? potential health benefits, savings, and costs. *Health Affairs*, 24(5):1103–1117, 2005.

[58] 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.

[59] David W Bates, Mark Ebell, Edward Gotlieb, John Zapp, and HC Mullins. A proposal for electronic medical records in us primary care. *Journal of the American Medical Informatics Association*, 10(1):1–10, 2003.

[60] Samuel J Wang, Blackford Middleton, Lisa A Prosser, Christiana G Bardon, Cynthia D Spurr, Patricia J Carchidi, Anne F Kittler, Robert C Goldszer, David G Fairchild, Andrew J Sussman, et al. A cost-benefit analysis of electronic medical records in primary care. *The American Journal of Medicine*, 114(5):397–403, 2003.

[61] Basit Chaudhry, Jerome Wang, Shinyi Wu, Margaret Maglione, Walter Mojica, Elizabeth Roth, Sally C Morton, and Paul G Shekelle. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine*, 144(10):742–752, 2006.

[62] Catherine M DesRoches, Eric G Campbell, Sowmya R Rao, Karen Donelan, Timothy G Ferris, Ashish Jha, Rainu Kaushal, Douglas E Levy, Sara Rosenbaum, Alexandra E Shields, et al. Electronic health records in ambulatory carea national survey of physicians. *New England Journal of Medicine*, 359(1):50–60, 2008.

[63] Cathy Schoen, Robin Osborn, Michelle M Doty, Meghan Bishop, Jordon Peugh, and Nandita Murukutla. Toward higher-performance health systems: adults health care experiences in seven countries, 2007. *Health Affairs*, 26(6):w717–w734, 2007.

[64] 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.

[65] 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.

[66] W Carl Cooley. Redefining primary pediatric care for children with special health care needs: the primary care medical home. *Current Opinion in Pediatrics*, 16(6):689–692, 2004.

[67] W Carl Cooley and Jeanne W McAllister. Building medical homes: improvement strategies in primary care for children with special health care needs. *Pediatrics*, 113(4):1499–1506, 2004.

[68] W Carl Cooley, Jeanne W McAllister, Kathleen Sherrieb, and Karen Kuhlthau. Improved outcomes associated with medical home implementation in pediatric primary care. *Pediatrics*, 124(1):358–364, 2009.

[69] Mark W Friedberg, Dana G Safran, Kathryn L Coltin, Marguerite Dresser, and Eric C Schneider. Readiness for the patient-centered medical home: structural capabilities of massachusetts primary care practices. *Journal of General Internal Medicine*, 24(2):162–169, 2009.

[70] Mark W Friedberg, Dana Gelb Safran, Kathryn Coltin, Marguerite Dresser, and Eric C Schneider. Paying for performance in primary care: potential impact on practices and disparities. *Health Affairs*, 29(5):926–932, 2010.

[71] 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.

[72] Toby Gosden, Frode Forland, Ivar Sonbo Kristiansen, Matthew Sutton, Brenda Leese, Antonio Giuffrida, Michelle Sergison, and Lone Pedersen. Impact of payment method on behaviour of primary care physicians: a systematic review. *Journal of Health Services Research & Policy*, 6(1):44–55, 2001.

[73] Stephen Campbell, David Reeves, Evangelos Kontopantelis, Elizabeth Middleton, Bonnie Sibbald, and Martin Roland. Quality of primary care in england with the introduction of pay for performance. *New England Journal of Medicine*, 357(2):181–190, 2007.

[74] 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.

[75] Diane R Rittenhouse, Stephen M Shortell, and Elliott S Fisher. Primary care and accountable caretwo essential elements of delivery-system reform. *New England Journal of Medicine*, 361(24):2301–2303, 2009.

[76] Mark Murray and Donald M Berwick. Advanced access: reducing waiting and delays in primary care. *The Journal of the American Medical Association*, 289(8):1035–1040, 2003.

[77] Mark Murray, Thomas Bodenheimer, Diane Rittenhouse, and Kevin Grumbach. Improving timely access to primary care: case studies of the advanced access model. *The Journal of the American Medical Association*, 289(8):1042–1046, 2003.

[78] Chris Salisbury. Does advanced access work for patients and practices? *British Journal of General Practice*, 54(502):330–331, 2004.

[79] Chris Salisbury, Stephen Goodall, Alan A Montgomery, D Mark Pickin, Sarah Edwards, Fiona Sampson, Lucy Simons, and Val Lattimer. Does advanced access improve access to primary health care? questionnaire survey of patients. *British Journal of General Practice*, 57(541):615–621, 2007.

[80] Francis G Belardi, Sam Weir, and Francis W Craig. A controlled trial of an advanced access appointment system in a residency family medicine center. *Family Medicine - Kansas City*, 36(5):341–345, 2004.

[81] Mark Pickin, Alicia O'Cathain, Fiona C Sampson, and Simon Dixon. Evaluation of advanced access in the national primary care collaborative. *British Journal of General Practice*, 54(502):334–340, 2004.

[82] Murat M Günal and Michael Pidd. Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4(1):42–51, 2010.

[83] Tillal Eldabi, RJ Paul, and T Young. Simulation modelling in healthcare: reviewing legacies and investigating futures. *Journal of the Operational Research Society*, 58(2):262–270, 2007.

[84] Jennifer L Wiler, Richard T Griffey, and Tava Olsen. Review of modeling approaches for emergency department patient flow and crowding research. *Academic Emergency Medicine*, 18(12):1371–1379, 2011.

[85] Catharine W Burt and Linda F McCaig. Trends in hospital emergency department utilization: United states, 1992-99. *Vital and Health Statistics. Series 13, Data from the National Health Survey*, (150):1–34, 2001.

[86] Stephen R Pitts, Richard W Niska, Jianmin Xu, and Catharine W Burt. National hospital ambulatory medical care survey: 2006 emergency department summary. *National Health Statistics Reports*, 7(7):1–38, 2008.

[87] Sharoda A Paul, Madhu C Reddy, and Christopher J DeFlitch. A systematic review of simulation studies investigating emergency department overcrowding. *Simulation*, 86(8-9):559–571, 2010.

[88] Geoffrey R Hung, Sandra R Whitehouse, Craig O'Neill, Andrew P Gray, and Niranjan Kissoon. Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. *Pediatric Emergency Care*, 23(1):5–10, 2007.

[89] Lloyd G Connelly and Aaron E Bair. Discrete event simulation of emergency department activity: A platform for system-level operations research. *Academic Emergency Medicine*, 11(11):1177–1185, 2004.

[90] David Sinreich and Yariv Marmor. Emergency department operations: the basis for developing a simulation tool. *IIE Transactions*, 37(3):233–245, 2005.

[91] Jennifer E Kreke, Andrew J Schaefer, and Mark S Roberts. Simulation and critical care modeling. *Current Opinion in Critical Care*, 10(5):395–398, 2004.

[92] Jeffrey Deacon Griffiths, M Jones, Martyn Sinclair Read, and Janet Elizabeth Williams. A simulation model of bed-occupancy in a critical care unit. *Journal of Simulation*, 4(1):52–59, 2010.

[93] Jeffrey Deacon Griffiths, Naomi Price-Lloyd, M Smithies, and Janet Elizabeth Williams. Modelling the requirement for supplementary nurses in an intensive care unit. *Journal of the Operational Research Society*, 56(2):126–133, 2005.

[94] Rodrigo B Ferreira, Fernando C Coelli, Wagner CA Pereira, and Renan MVR Almeida. Optimizing patient flow in a large hospital surgical centre by means of discrete-event computer simulation models. *Journal of Evaluation in Clinical Practice*, 14(6):1031–1037, 2008.

[95] Matthew Reynolds, Christos Vasilakis, Monsey McLeod, Nicholas Barber, Ann Mounsey, Sue Newton, Ann Jacklin, and Bryony Dean Franklin. Using discrete event simulation to design a more efficient hospital pharmacy for outpatients. *Health Care Management Science*, 14(3):223–236, 2011.

[96] Zexian Zeng, Xiaolei Xie, Xiang Zhong, Jingshan Li, Barbara A Liegel, and Sue Sanford-Ring. Simulation modeling of hospital discharge process. In *Proceedings of the 2013 IIE Annual Conference*, page 1383. Institute of Industrial Engineers-Publisher, 2013.

[97] Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: A review of literature. *Production and Operations Management*, 12(4):519, 2003.

[98] Tugba Cayirli, Emre Veral, and Harry Rosen. Designing appointment scheduling systems for ambulatory care services. *Health Care Management Science*, 9(1):47–58, 2006.

[99] Paul Robert Harper and HM Gamlin. Reduced outpatient waiting times with improved appointment scheduling: a simulation modelling approach. *OR Spectrum*, 25(2):207–222, 2003.

[100] S Noyan Ogulata, M Oya Cetik, Esra Koyuncu, and Melik Koyuncu. A simulation approach for scheduling patients in the department of radiation oncology. *Journal of Medical Systems*, 33(3):233–239, 2009.

[101] R Ashton, L Hague, M Brandreth, D Worthington, and S Cropper. A simulation-based study of a nhs walk-in centre. *Journal of the Operational Research Society*, 56(2):153–161, 2005.

[102] 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.

[103] Vanda De Angelis, Giovanni Felici, and Paolo Impelluso. Integrating simulation and optimisation in health care centre management. *European Journal of Operational Research*, 150(1):101–114, 2003.

[104] James E Stahl, Mark S Roberts, and Scott Gazelle. Optimizing management and financial performance of the teaching ambulatory care clinic. *Journal of General Internal Medicine*, 18(4):266–274, 2003.

[105] Thomas R Rohleder, Diane P Bischak, and Leland B Baskin. Modeling patient service centers with simulation and system dynamics. *Health Care Management Science*, 10(1):1–12, 2007.

[106] Christos Vasilakis, BG Sobolev, Lisa Kuramoto, and AR Levy. A simulation study of scheduling clinic appointments in surgical care: individual surgeon versus pooled lists. *Journal of the Operational Research Society*, 58(2):202–211, 2007.

[107] Diwakar Gupta and Brian Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE Transactions*, 40(9):800–819, 2008.

[108] Canan Pehlivan, Vincent Augusto, Xiaolan Xie, and Catherine Crenn-Hebert. Multi-period capacity planning for maternity facilities in a perinatal network: A queuing and optimization approach. In *2012 IEEE International Conference on Automation Science and Engineering*, pages 137–142. IEEE, 2012.

[109] Gregory Dobson, Hsiao-Hui Lee, and Edieal Pinker. A model of icu bumping. *Operations Research*, 58(6):1564–1576, 2010.

[110] Gad Abraham, Graham B Byrnes, Christopher Bain, et al. Short-term forecasting of emergency inpatient flow. *IEEE Transactions on Information Technology in Biomedicine*, 13(3):380–388, 2009.

[111] Maria Pia Fanti, Agostino Marcello Mangini, Mariagrazia Dotoli, and Walter Ukovich. A three-level strategy for the design and performance evaluation of hospital departments. *IEEE Transactions on Systems, Man, and Cybernetics*, 43(4):742–756, 2013.

[112] Vincent Augusto and Xiaolan Xie. A modeling and simulation framework for health care systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 44(1):30–46, 2014.

[113] Bo Zeng, Ayten Turkcan, Ji Lin, and Mark Lawley. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research*, 178(1):121–144, 2010.

[114] Wen-Ya Wang and Diwakar Gupta. Adaptive appointment systems with patient preferences. *Manufacturing & Service Operations Management*, 13(3):373–389, 2011.

[115] Christos Zacharias and Michael Pinedo. Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, 23(5):788–801, 2014.

[116] Victor Irvine, Sally McClean, and Peter Millard. Stochastic models for geriatric in-patient behaviour. *Mathematical Medicine and Biology*, 11(3):207–216, 1994.

[117] Sally McClean, B McAlea, and Peter Millard. Using a markov reward model to estimate spend-down costs for a geriatric department. *Journal of the Operational Research Society*, 49(10):1021–1025, 1998.

[118] Gordon Taylor, Sally McClean, and Peter Millard. Continuous-time markov models for geriatric patient behaviour. *Applied Stochastic Models and Data Analysis*, 13(3-4):315–323, 1997.

[119] Gordon Taylor, Sally McClean, and Peter Millard. Stochastic models of geriatric patient bed occupancy behaviour. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(1):39–48, 2000.

[120] 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.

[121] 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.

[122] Xiaolei Xie, Jingshan Li, Colleen H Swartz, and Paul DePriest. Improving response-time performance in acute care delivery: a systems approach. *IEEE Transactions on Automation Science and Engineering*, 11(4):1240–1249, 2014.

[123] Bernard S Bloom. Crossing the quality chasm: a new health system for the 21st century. *The Journal of the American Medical Association*, 287(5):646–647, 2002.

[124] Should physicians use email to communicate with patients? *Wall Street Journal*, 2012.

[125] Steven J Katz, Cheryl A Moyer, Douglas T Cox, and David T Stern. Effect of a triage-based e-mail system on clinic resource use and patient and physician satisfaction in primary care. *Journal of General Internal Medicine*, 18(9):736–744, 2003.

[126] Steven J Katz, Neil Nissan, and Cheryl A Moyer. Crossing the digital divide: evaluating online communication between patients and their providers. *The American Journal of Managed Care*, 10(9):593–598, 2004.

[127] Hessam Bavafa, Sergei Savin, and Christian Terwiesch. Managing office revisit intervals and patient panel sizes in primary care. *Available at SSRN 2363685*, 2013.

[128] Hessam Bavafa, Lorin M Hitt, and Christian Terwiesch. Patient portals in primary care: Impacts on patient health and physician productivity. *Available at SSRN 2363705*, 2013.

[129] Linda Green, Peter Kolesar, and Anthony Svoronos. Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research*, 39(3):502–511, 1991.

[130] Jingshan Li and Semyon M Meerkov. *Production Systems Engineering*. Springer, 20009.

[131] Jingshan Li and Semyon M Meerkov. On the coefficients of variation of uptime and downtime in manufacturing equipment. *Mathematical Problems in Engineering*, 2005(1):1–6, 2005.

[132] Xiang Zhong, Jie Song, Susan M Ertl, Jingshan Li, and Lauren Fielder. Design and analysis of gastroenterology (gi) clinic in digestive health center: A systems approach. *Flexible Service and Manufacturing*, 28(1):90–119, 2016.

[133] Alexander R Margulis and Jonathan H Sunshine. Radiology at the turn of the millennium 1. *Radiology*, 214(1):15–23, 2000.

[134] John A Buzacott and J George Shanthikumar. *Stochastic Models of Manufacturing Systems*, volume 4. Prentice Hall Englewood Cliffs, NJ, 1993.

[135] Yves Dallery and Stanley B Gershwin. Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems*, 12(1-2):3–94, 1992.

[136] Robert B Cooper. *Introduction to Queueing Theory*. Macmillan, 1972.

[137] Hideaki Takagi. *Queueing Analysis, Vol. 1: Vacation and Priority Systems*. 1991.