HOSPITAL RAPID RESPONSE SYSTEM: MODELING, ANALYSIS AND IMPROVEMENT

By

Xiaolei Xie

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Industrial Engineering)

at the

UNIVERSITY OF WISCONSIN – MADISON

2014

Date of final oral examination: 05/07/2014

The dissertation is approved by the following members of the Final Oral Committee:
Jingshan Li, Associate Professor, Industrial Engineering
Leyuan Shi, Professor, Industrial Engineering
Douglas Wiegmann, Associate Professor, Industrial Engineering
Neil Duffie, Professor, Mechanical Engineering
Marlon Mundt, Assistant Professor, Department of Family Medicine

© Copyright by Xiaolei Xie 2014

All Rights Reserved

To my parents, Lei Xie and Shouzhen Wu

Acknowledgements

I would first express my gratitude to my advisor Prof. Jingshan Li for his tremendous support throughout my graduate experience in Madison. He provides me with great research guidance that I can possibly think of. He also supports me meeting with other researchers in conferences, exploring intern opportunities and finding a job. His incredible mentorship has made a great deal of difference in my life. I would also express my gratitude to Prof. Leyuan Shi. Without her, I can never have this opportunity to join ISyE Dept. at UW-Madison, which is one of the best in US. Her caring, guidance and high standard have always propelled me to be a better scholars. I am extremely fortunate to meet them in my life. Also, I would thank my committee members, Prof. Douglas Wiegmann, Prof. Neil Duffie, Prof. Marlon Mundt. They are great scholars in their respective research and it is a precious opportunity to be challenged and learn from them.

I would thank Dr. Paul Depriest from Baptist Memorial Health System, Ms. Colleen Swartz from University of Kentucky Hospital for their collaborative work in hospital rapid response research. I would also thank Dr. Yue Dong, Dr. Thomas R. Rohleder and Dr. Jeanne Huddleston from Mayo Clinic for the summer research trainee experience. I would thank NSF with funding number CMMI-1063671 for my research support.

During my PhD study, I am lucky to have my colleagues Junwen Wang, Wei Feng, Feng Ju, Xiang Zhong, Cong Zhao, Zexian Zeng, Weiwei Chen, Tao Wu, Aaron Armstrong, Siyang Gao, Hongbo Meng and Xiufeng Shao, who provide help to my research. They are all friends to me and we have had a great time together, which has made my

stay in Madison enjoyable.

Lastly, I would thank my parents for their unconditional love and deep caring, which always give me strength facing challenges in my life. This dissertation is lovingly dedicated to them.

Abstract

The number of preventable death in US hospital is astonishingly high, upward of 100,000 per year. This has prompted a national initiative to create and deploy rapid response team (RRT), to quickly identify, evaluate, triage, and treat patients with clinical signs of deterioration to reduce the frequency and severity of negative outcomes. Although implementation of RRT does demonstrate some significant reduction in hospital mortality in the wards, such improvement may not be consistent in clinical outcome. Therefore, the evidence of the efficacy of RRTs is limited.

To ensure a successful implementation of rapid response operations, the early identification, better recognition, timely and appropriate treatment, and proper structural organization of care providers are of critical importance. In particular, prompt response and treatment play a key role. Therefore, improvement of the existing rapid response process to facilitate quick response and intervention is necessary and important. While RRT plays a critical role in this, it is not the whole picture of hospital rapid response. Other providers will also be activated and provide care to a declining patient. Therefore, collaborative and integrated operations involving all the care providers including RRT, considered as rapid response system (RRS), should also be examined. Although there exist clinical research in topics related to RRT and RRS, mathematical models to study rapid response operations from a system perspective can provide a fresh look, and more importantly an analytical insight at the rapid response process. Since such models do not exist in the current literature, the goal of this study is intended to contribute to this end by developing the models and applying it for improvement.

To achieve this goal, we consider two types of RRSs, one with a dedicated RRT or one with an assembled RRT in which providers come from other divisions (such as intensive care unit, ICU). In the former model, we focus on reducing the decision time (i.e., the time from detection of patient declining to the time a final decision is made) and its variability, while the latter emphasizes on studying multiple processes involved in rapid response operations, including triage, patient declining, floor intervention, RRT call initiation, and coordination with ICU.

More specifically, for RRS with a dedicated RRT, analytical formulas to evaluate the mean, the coefficient of variation, and the distribution of decision time are developed. System-theoretic properties are investigated. A bottleneck analysis method is introduced to identify and improve the process whose improvement can lead to the largest improvement in system performance. Indicators based on the data collected on the hospital floor are discovered to facilitate the identification and mitigation of bottleneck responses. Lastly, using a two-level recursive procedure, we consider multiple patient scenario and resource sharing during rapid response process is addressed. For RRS with an assembled RRT, an analytical framework is developed to study the correlations and coordination among different departments and multiple providers involved in RRS. With such a framework, a continuous time Markov chain model is introduced to evaluate the steady state probabilities of patient status. Analytical formulas are developed for the single patient scenario, and a recursive procedure is presented to study the multiple-patient case. Finally, a case study at University of Kentucky Chandler Hospital is introduced to illustrate the applicability of the method. It is shown that the developed model provides accurate estimation of system performance.

In summary, improving patient safety is the top priority for hospital management.

The models and methods developed in this study will provide a quantitative tool to analyze and improve rapid response operations from a system point of view.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Motivation

The expenditures in healthcare in US reached $2.7 trillion, which consumed about 17.9 percent of GDP in 2011. It is projected that such spending will be almost 20 percent by 2021, according to Centers for Medicare and Medicaid Services (National Health Statistics Group [53]). However, such a high cost does not make the healthcare delivery as safe as it should be. The number of preventable death is astonishingly high, upward of 100,000 per year, according to the report, "To Err is Human," released by Institute of Medicine (Kohn et al. [68]). Therefore, improving patient safety is always the top priority for hospital management. One of the hospital quality improvement efforts is to focus on identifying hospital deteriorating patients and reducing preventable harm and mortality. It has been shown that more than 80% of patients in hospital wards show signs of physiological deterioration before ICU admission (Goldhill et al. [47]). Up to 41% of ICU admissions are considered avoidable due to suboptimal care (McQuillan et al. [81]). Therefore, quick response and appropriate treatment to patient deterioration is critical (Downey et al. [38]), which has attracted nationwide interests in recent years to improve hospital care quality and ensure patient safety (Shojania et al. [97]). In particular, This has prompted a national concern (Anderson et al. [7], Brindley [19], Kaldjian et al. [66], Landrigan et al. [70], Leape and Berwick [71], Wachter [107], Winters et al. [116])

and initiative to create and deploy rapid response teams (RRTs, also known as medical emergency teams, METs) to quickly evaluate, triage, and treat patients with clinical signs of deterioration to against negative outcome (Berwick et al. [13], Hillman et al. [61], Priestley et al. [88]).

Clinical deterioration in the hospital setting is often a precursor to serious and often fatal outcomes, such as cardiac arrest and unplanned admissions to the intensive care unit (ICU). Intuitively, the implementation of RRT makes sense to provide a systematic response to deterioration episodes, and it does demonstrate significant reduction in hospital mortality in the wards in some cases, as observed in a study by Priestley et al. [88]. However, such improvement may not be consistent in clinical outcome (Winters et al. [115]). Therefore, the evidence of the efficacy of RRTs is limited.

To ensure a successful implementation of rapid response operations in acute care delivery, the early identification, better recognition, timely and appropriate treatment, and proper structural organization of care providers are of critical importance. In particular, the prompt response and treatment play a key role. Several studies indicate that although the deterioration information may be available, the response to such information remains a concern, and the activation of the responses may be problematic (Buist et al. [20], Franklin and Mathew [43], Goldhill et al. [46], Hillman et al. [60], Smith and Wood [102]). Therefore, adjustment and optimization of the existing rapid response model to facilitate quick response and intervention to improve outcome are necessary and important (Downey et al. [38]).

To achieve this, a mathematical model of the rapid response operations can provide a fresh look at the acute care delivery process. However, no such models are available in the current literature. In addition, improving the rapid response process to rescue patients is a system problem. Although RRT plays a central role, it also involves many

other providers from multiple disciplines and depends on divisions and structural configuration of workforce, protocol and operation strategies. Therefore, studying a rapid response system (RRS), rather than RRT only, is important. It is necessary to establish an analytical framework by developing quantitative engineering methods for analysis, design, and continuous improvement of RRS in acute care delivery for deteriorating patients and apply the results obtained on hospital floors to assist managerial decision-makings.

Hospitals across the nation may have different RRSs. In many teaching hospitals, RRT is a dedicated team consisting mainly of experienced nurses. In some hospitals, RRT is an assembled team often composed of various providers, such as fellow doctors from ICU. The advantage of the first type of RRT is its generally shorter activation time which better ensures "quick" response. But, for patients going through complicated deterioration, further help from more experienced providers is often needed. Therefore, in the study of RRS with a dedicated RRT, we focus on timely treatment to patients with clinical deteriorations. The second type of RRT may have more experience and usually is capable of making final medical decision. However, the time of assembling the team and traveling from another division is longer, as the physicians in the RRT need to hold their current work in their own department, such as ICU, and then respond to RRT calls. In some hospitals, they may need to travel from different buildings. Therefore, for RRS with an assembled RRT, we investigate the entire rescuing process and coordinations between different divisions with an emphasis on resource utilization study. The following two subsections address the research motivation in the dedicated and assembled RRT scenarios, respectively.

1.1.1 RRS with Dedicated RRT

It has been shown that patient safety and care quality are strongly correlated with care delivery time for a declining patient (Franklin and Mathew [43] and Hillman et al. [60]). Thus, the activation of the response and timely call for assistance are critical to ensure quick delivery of care and appropriate treatment (Downey et al. [38]). Therefore, in this study, we focus on reduction of the decision time (i.e., from the time when a declining signal is detected to the time a final treatment decision, such as admission to ICU, is made) to improve patient safety and care quality. In practice, in addition to the average response time, its variability may also have a strong influence on the patient outcome. Clearly, large variance may result in severe negative outcomes. Hence, reducing the variability is also of tremendous importance. However, variance itself does not directly specify how large the variability is without the knowledge of the average decision time. Thus, we focus on the coefficient of variation (CV) in reducing decision time variability.

Although variance or CV can provide a general and theoretical characterization of the variability, it does not specify whether a desired care delivery time is satisfied or not. Therefore, a more direct measure, the response-time performance (RTP), which is referred to as the probability to make a final decision within a desired time period, is more preferable in practice. In acute care, the available time to save a patient is limited, which results in short desired time intervals. Thus, the RTP cannot be obtained using Central Limit Theorem. A direct method for evaluating RTP is needed.

Using the analytical model developed, system properties, such as monotonicity, can be investigated. Understanding these properties can provide us the direction for improving the efficacy of RRS. Following such directions, one needs to know which response process impedes the system performance in the strongest manner. In other words, improving which one can lead to the largest improvement in RRS? Such a response is

referred to as the bottleneck response. However, identifying such a response is not easy, since it needs to evaluate all the response performance and the impacts of this variations. Therefore, a bottleneck analysis method based on the data collected on the hospital floor is needed to identify the most critical process (i.e., bottleneck) in the response system, so that the decision-makers can focus on for continuous improvement purposes. Lastly, we extend the study and consider the cases where multiple patients requesting one care provider, which leads to extra waiting. A two-level iterative procedure is developed to evaluate the new average decision time.

Chapter 3 is focused on developing such models, investigating the system properties, and discovering the bottleneck indicators for identification and improvement of bottleneck responses and finally the extension.

1.1.2 RRS with Assembled RRT

RRT intervention is part of the hospital rescuing effort, which is activated after acute physiological deterioration happens. Ultimately, the goal is to improve hospital rapid response to prevent undesirable outcomes. In the case of assembled RRT, the providers in RRT typically come from ICU. Coordination between RRT activation and ICU operations becomes important. A system model can provide a new perspective of the problem. Through quantitative modeling of the rescue process and studying the issues related to operation protocols, RRT strategies, and team composition, etc., recommendations to improve the rescue process can be provided.

To achieve this, an analytical framework, which consists of the modules related to RRS, is established to characterize the patient rescue process. The modules include:

- A triage module assigns patients to floor ward or ICU.

- A patient module provides patient status: normal or deteriorating.
- A floor module describes patient monitoring and provider (nurse/RN and doctor/MD) intervention.
- A RRT module represents RRT activation and treatment.
- An ICU module involves RRT composition and response.

Clearly, all these modules are involved in RRS and they are interacting with each other. For example, RRT treatment is critical to bring patient back to normal status, while frequent RRT calls may impact providers' work in ICU. Therefore, using this framework, an analytical model is needed to characterize the patient rescuing states and interactions among different modules (i.e., patient, multiple providers, and departments). Such a model can provide insights to aid decision making on staffing allocation, team composition, detection, intervention, and RRT call protocols, etc.

Chapter 4 introduces the analytical framework and a continuous time Markov chain model to estimate the performance of RRS with an assembled RRT.

1.2 Organization of the Document

The rest of this document is organized as follows. Chapter 2 reviews the related literature in both clinical and engineering areas. Chapter 3 presents the analytical model of RRS with dedicated RRTs. Formulas for performance evaluation are derived. System properties and continuous improvement methods are also examined. A iterative procedure is developed the address an extension of the current problem. In Chapter 4, the five-module analytical framework for patient rescue process with assembled RRTs is introduced and a continuous time Markov chain model is presented. Analytical formula

and iteration procedure are developed for performance evaluation. A case study is provided in Chapter 5 to demonstrate the applicability of the methods developed. All the proofs are provided in the Appendix.

Chapter 2

Literature Review

The goal of this research is to develop a rigorous engineering methodology to improve rapid response in hospitals. The following literature review focuses on both the studies through clinical trial approaches, discussed in Section 2.1, and using Industrial Engineering and Operations Research (IEOR) methods, introduced in Section 2.2. Section 2.3 summarizes the existing research and the perspective of the current study.

2.1 Clinical Trial Approach

Enhancing patient safety is the most important goal in hospital management. There has been a nationwide interest in recent years to study how to improve hospital care quality, particularly focusing on controlling patient mortality rate. Consequently, the concept of failure-to-rescue (FTR) is raised, which is a performance measure on inability to save a hospitalized patient's life when he or she experiences a complication, (Silber et al. [98]). Taenzer et al. [104] provide a comprehensive review of approaches to address FTR: Retrospective analysis, risk scoring system, monitoring system, effect of medical emergency team, etc., are examined. They conclude that continuous patient monitoring should be the next step for early intervention in the FTR field. Schmid et al. [93] also provide a review to enhance understanding of FTR from nurses' standpoint. Tested interventions are examined and implications for future research are presented. Clarke

[27] also discusses FTR from the perspective of nurse researcher and healthcare managers and emphasizes that a favorable condition should be created for nurse practitioners to assume individual accountability.

To reduce FTR, RRTs have been created and implemented to quickly evaluate, triage and treat patients with physiological deterioration. Deployment of such teams is discussed by Berwick et al. [13]. It is claimed that RRT can save lives either by initiating changes in care or by facilitating transfer to ICU. Priestley et al. [88] discuss a generalized idea of implementing RRT, which is termed as critical care outreach service because many RRTs are composed of critical care fellows, and conclude that such outreach reduces mortality in general hospital wards. The Institute for Healthcare Improvement has identified the deployment of RRTs as one of the major changes to prevent death in patients who are progressively experiencing acute physiological deterioration outside the ICU (IHI [1]). To study the effect of RRT on healthcare outcomes, Jones et al. [64] provide a summary of studies of RRTs involving comparison data, where all the findings suggest a positive correlation between RRT implementation and serious adverse events reduction, such as cardiac arrest. Through case studies in hospitals, Sharek et al. [96] show that, in a 264-bed children's hospital, RRT implementation is associated with a statistically significant reduction in hospital-wide mortality rate and code rate outside of the pediatric ICU setting. In another study, using the results in 350-bed community hospital, Dacey et al. [32] suggest that RRT deployment is associated with significant decreases in rates of in-hospital cardiac arrest and unplanned ICU admissions. Through the measurement of incidence of adverse events after major surgery, Bellomo et al. [12] conclude that the introduction of MET/RRT was associated with a reduced postoperative adverse outcomes, mortality rate and average hospital stay.

Although a number of successful cases are observed in different hospitals, undesirable

results after RRT deployment are also observed. Hillman et al. [61] design a prospective cluster-randomized controlled trial and conclude that MET/RRT does not substantially affect the incidence of cardiac arrest, unplanned ICU admissions, or unexpected death. Through the study in a 404-bed Kansas City-based hospital, Chan et al. [24] conclude that RRT implementation is not associated with reductions in hospital-wide code rates or mortality. Out of nine case studies investigated by Massey et al. [77], the results show that three of them see no positive impact on patient outcomes. Moreover, nurses are reluctant to use RRT with unclear rationale. Based on the review and meta-analysis of the studies published from year 1950 to 2008, Chan et al. [25] find out that there is a lacking in robust evidence to support effectiveness in reducing hospital mortality. In another review, Winters et al. [115] find out that there is a weak evidence between RRT's effective interventions and reduction of hospital mortality, cardiac arrest rate, or ICU admissions, due to the presence of biased results and other limiting factors. In addition, Litvak and Pronovost [75] re-examine the functionality of RRT and claim that many RRT activations are generated due to other factors in the hospital, such as triage error or limited resource. The unnecessary RRT calls can be eliminated through better patient flow management.

To more effectively evaluate RRT, the whole RRS needs to be studied, in which RRT acts as a key component. The impact of critical care outreach services on hospital mortality rates is studied by McGaughey et al. [79] and the results show no reduction in mortality rate. Ranji et al. [89] evaluate the effects of RRSs on clinical outcomes through a systematic literature review. Consistent improvement in clinical outcomes after deployment of RRS is not found in the review. DeVita et al. [35] introduce a RRS structure. It has an afferent limb, which is about event detection and triggering and efferent limb, which contains RRT intervention and further actions. They also provide

outcome measures and obstacles to implement RRS. DeVita et al. [36] further claim that most of the studies focus on the efferent limb, while afferent limb was overlooked. They suggest implementation of much more effective monitoring system. Trinkle and Flabouris [106] introduce afferent limb failure (ALF), and by analyzing retrospective medical record and database review, they conclude that ALF, as useful performance measure for RRS, is associated with unanticipated ICU admissions, and that the duration of ALF is associated with hospital mortality. Galhotra et al. [45] conduct a retrospective observational study in a 730-bed hospital. They conclude that more frequent monitoring and timely RRT activation are associated with acute care delivery improvement, while additional number of RRT calls might not prevent undesirable outcomes. Sebat et al. [94] study changes of times to key interventions and mortality rate during a 7 year period and conclude that after implementation of rapid response system, those measurement reduces significantly. Lastly, Peberdy et al. [86] recommend core dataset and measures which can guide hospitals to collect the most meaningful data to optimize RRS interventions.

In summary, through clinical trial approach, topics in FTR, RRT and RRS are investigated. To reduce FTR, RRT, which is a crucial indicator of hospital's capability to ensure patient safety, is introduced throughout the nation. However, consistent improvement results after RRT implementation are lacking. An integrated study to address RRT within RRS framework is necessary. Therefore, to further investigate RRT/RRS, a fresh look from quantitative model point of view using IEOR methods is needed.

2.2 IEOR Methods

Due to the complexity in RRS, a system model to quantify the correlations among various factors in RRS is important, which can provide a fresh look from another perspective.

IEOR techniques play a key role in such studies. Over the past years, methodologies in IEOR have tried to contribute to improve the quality of health care delivery. Brandeau et al. [17] provide a list of problems in health care area that can be solved by IEOR techniques. Problems such as capacity planning, resource scheduling, treatment plan improvement, patient flow management, etc., have been studied intensively (see reviews by Fomundam and Herrmann [42], Gupta and Denton [54], Cardoen et al. [21], Wiler et al. [114], and representative papers from Green [49], Dobson et al. [37], Wang et al. [109], Helm et al. [57], and Wang et al. [111]). For capacity planning problems, Green [50] describes the related background and issues, and provides examples of OR models developed to provide insights to operational strategies and practices. Harper and Shahani [56] present a detailed simulation model to help the planning and management of hospital beds, which enables hospital decision-makers to gain insights of the consequences of planning and management policies. Zhang et al. [121] introduce a simulation optimization approach to address hospital long term capacity planning problem through two case studies. For scheduling problems in healthcare, Cayirli and Veral [23] review the literature related to outpatient scheduling. Another comprehensive literature review of operating room scheduling is conducted by Cardoen et al. [21]. Denton et al. [34] develop a stochastic optimization model and present some heuristics to create operating room schedules under time uncertainty of surgery duration. Muthuraman and Lawley [83] introduce a stochastic overbooking model and develop an appointment scheduling policy for outpatient clinics, whose performance is investigated under different conditions. Beaulieu et al. [10] present a mathematical programming approach for preparation of physician schedule in the emergency room, which significantly reduce the time and effort to construct the schedule. Medical decision making in healthcare, such as radiation treatment planning, has attracted substantial amount of research interests. Brahme [16]

advocates advanced treatment techniques to improve the efficacy of radiation therapy and predicts that new powerful tools will be introduced. D'Souza et al. [39] present a nested partitions framework that helps find suitable beam angle sets by guiding the dose optimization process. Zhang et al. [120] develop a two stage solution approach to provide guidelines for physicians on beam angle selection and dose amount usage, which lead to a significant improvement in solution quality and result delivery time.

Patient flow is another area where significant amount of research has been developed. Both analytical methods, such as queueing and Markov chain models, and simulation approach are developed to address issues related to patient flow. Haraden and Resar [55] introduce methods and generate discussion for solving various problems on patient flow in order to reduce delays during care delivery. Fomundam and Herrmann [42] summarize the research work in waiting time reduction, utilization improvement, system design and appointment systems, using queueing models. Wiler et al. [114] compare different modeling approaches, such as regression models, time-series analysis, queuing models, and simulation. For each of the methodologies, they describe the fundamental assumptions and outline the potential applications and limitations. Green [48] describes basic queueing models and their extensions that are particularly useful in healthcare setting. As to specific applications, Weiss et al. [112] introduce a continuous time semi-Markov model of population flow within a network of service facilities and use such a model to predict length-of-stay distributions of a university teaching hospital. Wang et al. [109] introduce a Markov chain model to analyze the work flow in CT process and investigate the impact of staffing level at the imaging center of University of Wisconsin Medical Foundation. Yankovic and Green [117] introduce a finite source queueing model to analysis the dynamics of bed occupancy level and nursing demand, which help cost-effective staffing decision-making. Mandelbaum et al. [76] address issues in patient routing from

emergency department and hospital internal wards by developing a queueing system model with heterogeneous server pools and compare different routing policies.

In addition to analytical models, discrete event simulation (DES) is also a prevailing tool due to its flexibility to adjust many dynamic health care delivery settings. Jun et al. [65] provide a comprehensive survey of DES applied to health care systems, specifically in topics of patient flow and resource allocation. They further suggest the development of comprehensive simulation modeling framework for determining clinical performance measures and interdepartmental resource relationships. Jacobson et al. [62] and Taylor et al. [105] also present reviews of DES applied to healthcare. Substantial amount of simulation studies have been devoted to different topics, including improvement of a variety of hospital departments, such as emergency, pharmacy, primary care clinic etc; analysis of hospital capacity; staffing and work schedule determination and recommendations, and multidisciplinary work, such as discharge process, etc. In application studies, a simulation study in the emergency department at the University of Kentucky Chandler Hospital is conducted by Brenner et al. [18]. It provides a quantitative tool for continuous improvement and process control in the emergency department and the recommendation of adding a CT scanner and additional nurses has been implemented in hospital. In another application, Zeng et al. [119] aim at improving the quality of care at the emergency department of a community hospital in Lexington, Kentucky, and a DES model is developed. The suggestion of adding equipment and implementing limited team nursing policy is recommended to the hospital management. Coelli et al. [29] conduct the analysis of a mammography clinic performance using simulation and conclude that the actual capacity of the clinic is over-sized for its present demand. Zhong et al. [122] develop a simulation model to study the patient flow in pediatric care processes in the University of Wisconsin Health Systems, which help reduce unnecessary variation

in care processes, ensure efficiency of staff and engage patients in process improvement. Spry and Lawley [103] develop a simulation model to evaluate the effect of changes to staffing and work scheduling, which aims at helping BroMenn, a healthcare organization, find the best schedule to get medications to the patients as quickly as possible while using pharmacy staff effectively. Fung Kon Jin et al. [44] identify the best performing patient flow management strategy on the institutional level on process quality through simulation. Crawford et al. [30] use simulation model to help determine the appropriate time to discharge acute care patients. They study the impact of individual inpatient discharge decisions on other operations, such as ED crowding and readmission. Zeng et al. [118] also develop a simulation model which can accurately emulate the discharge process in University of Wisconsin Hospital and clinics and successfully identify system bottleneck. To sum up, IEOR methods, including both analytical approaches and DES, have attempted to address a variety of problems faced by healthcare systems. Results and insights from the developed models can be used to assist hospital managerial decision-makings.

Due to the similarities between healthcare delivery system and production systems, applying production systems engineering (PSE) methods in healthcare systems arises naturally. Using production systems research methodology, complex healthcare systems can be represented by structures similar to production networks. Brenner et al. [18], Wang et al. [108] and Zeng et al. [119] introduce a complex network with split, merge, parallel, closed and reentrant structure to model patient flow in emergency department. Wang et al. [110] and Wang et al. [111] transform the workflow of care delivery service within patient room from a closed and reentrant process into serial ones. Using such networks, improvement strategies can be developed.

Bottleneck identification and elimination have been viewed as the most effective way

to improve system performance (see monograph, Li and Meerkov [74], and representative studies by Biller et al. [14], Chiang et al. [26], Kuo et al. [69] and Li [73]). Applying bottleneck analysis in healthcare is also useful for detecting the most critical process for operational improvement. Brenner et al. [18] identify CT scanning as the bottleneck in emergency department. Similar results are obtained in Zeng et al. [119]. In addition, Wang et al. [109] discover the image review process as the bottleneck in CT work flow. Shao et al. [95] introduce bottleneck analysis to discover the most critical disruptions whose reduction can ensure most normal status in surgery.

In summary, although substantial research has been devoted to healthcare systems, the study on rapid response systems using mathematical models is still missing. Developing methods to address RRS, such as performance evaluation and continuous improvement, is of significant importance.

2.3 Perspectives

Over the years, it is undeniable that IEOR techniques intends to help improve the health care operations. However, regarding the topic to ensure patient safety, which is among the most critical issues in healthcare, there is a significant lacking in the contribution from IEOR literature. More specifically, although substantial studies have been carried out to study the effectiveness of RRT efforts, most of them are clinical trial based (DeVita et al. [35], Trinkle and Flabouris [106]). Few research is devoted to studying the issues from a system point of view. Similarly, studies on "failure to rescue" are also clinical based and have been addressed mostly from the point of view of nurse staffing level (Silber et al. [99], Needleman et al. [84], Silber et al. [100], Schmid et al. [93], Clarke and Aiken [28], Kendall-Gallagher et al. [67]), rather than from system perspectives. The

goal of this study is intended to bridge this gap.

The major challenge of this study is that, many of the existing IEOR methodologies can not be directly applied to the study of hospital rapid response. For instance, mathematical programming is a powerful tool and sees its application in many deterministic settings to address healthcare problems, particularly in issues such as planning and scheduling (Oddoye et al. [85], Herring and Herrmann [58], Herring and Herrmann [59], Epstein et al. [41], Earnshaw et al. [40], Reynolds et al. [90]), medical decision making (e.g., radiation therapy and cancer screening etc., see Meyer et al. [82], Lee et al. [72], Romeijn et al. [92], Alagoz et al. [6], Ayer et al. [8]), where objectives and constraints can be clearly defined and interrelated by introduced variables. However, to address hospital rapid response, which is a complex random process involving patients and a variety of resources, deterministic methods are not suitable approach . Therefore, stochastic systems approaches with more applications are more appropriate.

One direct approach is discrete event simulation, which can mimic the true rapid response system well due to its flexibility in developing the model with much details. Only one simulation study related to rapid response is discovered, which is conducted by Carter and Blake [22] in the acute care setting. However, their emphasis is on reducing acute patient waiting and resource allocation, rather than directly tackling the issues raised during the rapid response. The disadvantage of simulation is that it is often case study based and typically suffers from long model development and simulation times, and requires details of data inputs. Most importantly, when compared to quantitative analytical methods, simulation can only provide results under specific settings, yet lack the capability to uncover underlying system nature. The interpretation of obtained results may be challenging, because it is hard to determine whether these outputs are results of system dynamics and interrelationship, or due to pure randomness (Banks et al.

[9]). Therefore, simulation is not favorable in establishing a framework to systematically study hospital rapid response.

In another direction, queue models have also seen their application in healthcare settings to analyze patient flow in emergency and other departments (Green et al. [52], Jiang and Giachetti [63], de Bruin et al. [33]). However, the applicability of queueing theory model in RRS problem settings has several major issues. First, hospital rapid response differs from many other queueing applications due to limited population in the system. The issues like queue length, heavy traffic, etc., are not the main interests, while decision time and its variability are the main factors. Second, the definition of "server" can be extremely difficult. For example, for rapid response system with dedicated RRT, RRT can be solely activated, but can be also requested to couple with another type of provider. Moreover, such resources are shared by multiple patients. Third, many queue models are restricted to some specific assumptions and computational intensity for complex cases. . Therefore, queueing modeling is not the most appropriate approach in studying hospital rapid response.

Lastly, there is research conducted using Markov decision model or its variation to tackle healthcare problems (Silverstein et al. [101], Welton and Ades [113], Begun et al. [11], Ades and Cliffe [2], Pinker and Tezcan [87]). In particular, Markov decision process (MDP) has many real world applications in healthcare policy-related problems (Alagoz et al. [5], Alagoz et al. [4], Green et al. [51]). For hospital rapid response, an appropriate value function of MDP is extremely tricky to determine, so directly applying MDP is not feasible.

In summary, innovative stochastic modeling approaches should be developed to model rapid response operations, analyze performance measures, identify and improve the most impeding process. In the subsequent chapters, analytical models are developed

to study the decision time and its variability in rapid response process with a dedicated RRT, and the correlations in patient rescue process with an assembled RRT. First, the process is represented using a network structure with split and merge. Then, for simple systems, through analysis of stochastic equations characterizing the process at hand, analytical solutions are derived. Using iterative procedures, larger systems can be analyzed. Moreover, empirical formulas are developed to analyze non-Markovian scenarios on the hospital floor. In addition to performance evaluation, system properties are investigated and bottleneck analysis is carried out by developing indicators using the data that can be collected on the hospital floor. Finally, the methods have been applied in acute care delivery process in the hospital. The development of such models and methods, and the analysis of rapid response systems can provide hospital management and health care professionals a quantitative tool to improve patient safety and quality of care.

Chapter 3

Rapid Response Process Modeling

3.1 Introduction

Since the current literature is lacking both methods for evaluating performance measures and quantitative tools for improvement of the rapid response operations, we develop a novel approach to model and analyze the rapid response operations, and to improve its efficacy. To achieve this, in this chapter, a complex network with parallel, split, and merge structures has been introduced to model the rapid response process. First, performance evaluation problems are studied. Analytical formulas have been derived to evaluate the mean and coefficient of variation (CV) of decision time, and response-time performance, which characterize the efficacy of rapid response operations. Then, using these formulas, we can investigate the methods to facilitate quick response and intervention to improve outcomes. More specifically, continuous improvement activities can be carried out by identifying the bottleneck responses so that its mitigation will lead to the largest improvement (reduction of decision time and its CV or increase of RTP). Specifically, the bottleneck response for decision time (BN-τ) and its variability (BN-cv), and response-time performance bottleneck ($BN - rtp$) are introduced, and bottleneck indicators based on the data collected on the hospital floor are developed. Finally, we consider an extension under our current framework, where potential waiting due to resource sharing in the rapid response process is studied. We develop a two-level iteration

approach to evaluate the new mean decision time. Satisfactory accuracy is obtained and the applicability of the approach is discussed. Our work in rapid response process modeling establish an analytical framework which provide a quantitative decision-making tool for hospital practitioners to improve rapid response system with dedicated RRT.

3.2 System Description and Problem Formulation

Consider a rapid response process in acute care delivery summarized as follows (see Figure 3.1 on next page):

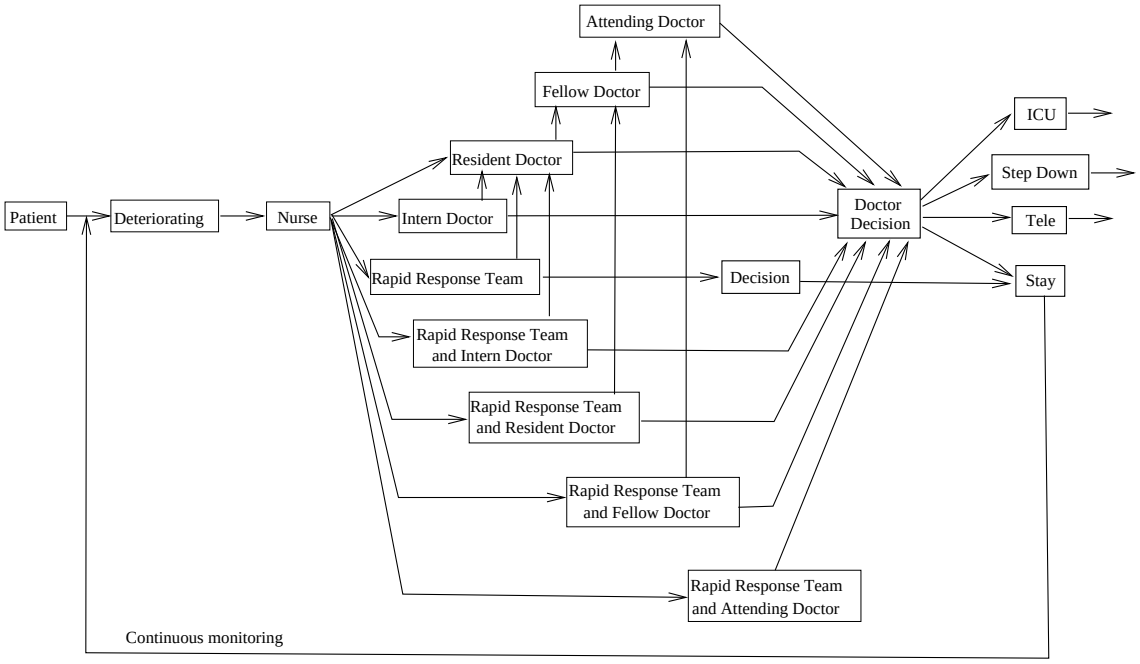

Figure 3.1: Rapid response process for declining patient in acute care

Procedure 3.1 Rapid response process flow:

- A patient in acute care is continuously monitored for vital signs, such as blood pressure, heart beat, etc. When a declining signal is identified, the primary nurse

should response quickly and ask for help. She may call RRT or an intern doctor, or both of them. If the patient's condition is complicated, she may also call the resident doctor directly, or both RRT and resident (sometimes even RRT and fellow, or RRT and attending).

- Upon receiving the call from the nurse, the RRT needs to arrive immediately and appropriate care should be carried out. Based on patient's condition, a decision on either keeping the patient stay for more observations or calling the resident doctor will be made.

- If the intern doctor is called for help, he (or she) needs to initiate appropriate treatment and decide whether a higher level (resident) doctor is needed or not. Otherwise one of the following four decisions: ICU, step down, telemetry, or s-tay, will be made, where "ICU" implies admission to intensive care unit, "tele" (telemetry) refers to moving from acute care to a monitored bed (where physiologic monitor presents), "step down" represents progressive care – a level higher than telemetry, but not as high as ICU, and "stay" stands for continuing observation. Similar case occurs when both RRT and intern doctor are showing up.

- Analogously, such logic applies to resident and fellow doctors (or both RRT and the doctors) as well. When they receive a help request (from nurse, or intern, or RRT), their diagnosis and treatment of the patient will lead to a decision of either calling to a higher level doctor (fellow or attending, respectively), or selecting one of the four possible decisions.

- The attending doctor (or attending and RRT) will make a final decision (to one of the four possible routes) if he is called.

Such a rapid response process can be viewed as a complex network with split, merge and parallel structures. Studying such a network to improve the efficacy of rapid response operations is the goal of this chapter. The assumptions formulated below define the patient, care providers, and decision flow.

Patient:

(i) In each patient room, there is one patient who may exhibit clinical deterioration.

Care providers:

(ii) The sets of care providers are defined as $X_1 = \{$nurse, RRT, intern, resident, fellow, attending$\}$ when single provider carries out the service, and $X_2 = \{$RRT & intern, RRT & resident, RRT & fellow, RRT & attending$\}$ when two providers work jointly.

(iii) The response process of care providers to diagnose and treat the declining patient is modeled as a "machine" (or "server") with random service time t_x, defined by the probability density function $f_x(t)$, where $x \in X = X_1 \cup X_2$. (Note that in the subsequent content, x is referred to as the response process or service process interchangeably.)

Decision flow:

(iv) The rapid response process follows the steps described in Procedure 3.1 and Figure 3.1.

(v) $Y = \{$ICU, stay, tele, step down$\}$ define the set of four possible final decisions the doctors (or doctors & RRT) can make. In a few cases, the rapid response team can make a stay decision.

(vi) The provider's choice of calling for help is characterized by probability α_{ij}, where $j \in X$ ($j \neq$ attending, or RRT & attending), and i is the upper level provider, as illustrated in Procedure 3.1 and Figure 3.1.

(vii) The doctor's final decision (to one of the four possible routes) is characterized by probability β_y^j, $y \in Y$, and $j \in X$, $j \neq$ nurse, RRT.

Remark 3.1 The assumptions introduced above represent a preliminary model of rapid response process in acute care delivery. Such a model can be extended to include more complex nature of the process. For instance, patients can be grouped based on their syndromes, and characterized by the states (or stages) of declining. The routing probabilities α and β will be functions of these groups, states, and availabilities of care providers. The decision quality and the impacts of the education and experience levels of care providers also need to be investigated.

In an appropriately defined state space, system (i)-(vii) is a stationary random process. Let t_d denote the decision time (from declining to final decision making) and T_d as its mean. In the framework of (i)-(vii), T_d is a function of all system parameters.

$$T_d = F_{t_d}(\mathcal{F}, \mathcal{A}), \tag{3.1}$$

where

$$\begin{aligned}
\mathcal{F} &= [f_n, f_{rrt}, f_{int}, f_{res}, f_f, f_a, f_{rrt\&int}, f_{rrt\&res}, f_{res\&f}, f_{res\&a}], \\
\mathcal{A} &= [\alpha_{int,n}, \alpha_{rrt,n}, \alpha_{res,n}, \alpha_{rrt\&int,n}, \alpha_{rrt\&res,n}, \alpha_{rrt\&f,n}, \alpha_{rrt\&a,n}, \alpha_{res,int}, \alpha_{res,rrt\&int}, \\
&\quad \alpha_{res,rrt}, \alpha_{f,res}, \alpha_{f,rrt\&res}, \alpha_{a,rrt\&f}, \alpha_{a,f}],
\end{aligned}$$

and subscripts n, rrt, int, res, f, a, $rrt\&int$, $rrt\&res$, $rrt\&f$, and $rrt\&a$ represent nurse, RRT, intern, resident, fellow, attending, RRT & intern, RRT & resident, RRT & fellow, and RRT & attending, respectively.

Since variations in rapid response operations play a critical role in system responsiveness, the variability of decision time, characterized by CV, is another important performance index. Therefore, we introduce the CV of decision time, CV_d, as follows:

$$CV_d = F_{cv}(\mathcal{F}, \mathcal{A}). \tag{3.2}$$

In order to determine whether a desired care delivery time is satisfied or not, the *response-time performance*, RTP, which is referred to as the probability that the decision time is less than a desired time period T_s, is introduced:

$$RTP = \text{Prob}(t_d \leq T_s), \tag{3.3}$$

where T_s could be syndrome dependent. Clearly, RTP is also a function of all system parameters for a given T_s, i.e.,

$$RTP = F_{rtp}(\mathcal{F}, \mathcal{A}, T_s). \tag{3.4}$$

We seek analytical characterization of T_d, CV_d, and RTP. Therefore, the problem to be addressed is:

Given the rapid response process (i)-(vii), develop a method to calculate the mean and coefficient of variation of decision time and response-time performance as functions of system parameters, and investigate system structural properties.

Remark 3.2 In addition to T_d and CV_d, the means and variations of decision time specific to the four possible destinations, step down, ICU, tele, and stay, can be obtained similarly by taking into account the probabilities β_y^j, $y \in Y$, and $j \in X$, $j \neq$ nurse, RRT.

$$\begin{aligned} T_{d,y} &= F_{t_d,y}(\mathcal{F}, \mathcal{A}, \mathcal{B}, Y), \\ CV_{d,y} &= F_{cv,y}(\mathcal{F}, \mathcal{A}, \mathcal{B}, Y), \\ RTP_y &= F_{rtp,y}(\mathcal{F}, \mathcal{A}, \mathcal{B}, T_s, Y), \end{aligned}$$

where

$$\mathcal{B} = [\beta^{int}_{ICU}, \beta^{int}_{stay}, \beta^{int}_{tele}, \beta^{int}_{stepdown}; \beta^{res}_{ICU}, \beta^{res}_{stay}, \beta^{res}_{tele}, \beta^{res}_{stepdown}; \ldots; \beta^{a}_{ICU}, \beta^{a}_{stay}, \beta^{a}_{tele}, \beta^{a}_{stepdown}],$$

$$Y = \{ICU, stay, tele, stepdown\}.$$

3.3 Performance Evaluation

The goal of this work is to evaluate the mean and coefficient of variation of decision time as well as response-time performance in responding to declining signals for the rapid response process defined by assumptions (i)-(vii). In other words, we would like to develop a technique for calculating T_d, CV_d, and RTP as functions of $\mathcal{F}$ and $\mathcal{A}$ (and $\mathcal{B}$ and Y). Such a technique can help hospital management predict the outcome for various values of system parameters and seek improvement strategies.

3.3.1 Decision Time and its Variability

From Figure 3.1, let l_{ij} denote the routing indicator from response j to response i. Then $l_{ij} = 1$ if the patient is routed from response j to response i, and $l_{ij} = 0$ if such a route is not selected. Then, define A_i as that response i has been initiated and carried out, $i \in X$, we obtain

$$A_i = \begin{cases} l_{i,n}, & \text{if } i \in \{int, rrt, rrt\&int, rrt\&res, rrt\&f, \\ & \quad rrt\&a\}, \\ l_{res,n} + \sum_{j=int,rrt,r\&i} l_{res,j} A_j, & \text{if } i = res, \\ l_{f,res} A_{res} + l_{f,rrt\&res} A_{rrt\&res}, & \text{if } i = f, \\ l_{a,f} A_f + l_{a,rrt\&f} A_{rrt\&f}, & \text{if } i = a. \end{cases} \tag{3.5}$$

Using the above notations, let $A_i = l_{i,n}$, $i \in \{int, rrt, rrt\&int, rrt\&res, rrt\&f, rrt\&a\}$, denote that the first response i to the nurse's call is carried out. The next service carried out by the resident doctor is characterized by A_{res}, where patients are routed from the nurse directly or from the first response (intern doctor, RRT, or RRT and intern doctor). A_f describes the fellow doctor's responses to the patients who finish the first (RRT & resident doctor) and second (resident doctor) ones but still need help. Finally, A_a represents the final response by attending doctor, working on the cases coming from RRT & fellow doctor and fellow doctor.

Therefore, $A_i = 1$ implies that service i has been carried out for the patient, while $A_i = 0$ suggests that the patient does not go through this service. Let p_i denote the probability that response i, $i \in X$, has been carried out, i.e.,

$$p_i = \text{Prob}(A_i = 1).$$

Then, we have

$$p_i = \begin{cases} 1, & \text{if } i = n, \\ \alpha_{i,n}, & \text{if } i \in \{int, rrt, rrt\&int, rrt\&res, rrt\&f, \\ & \quad rrt\&a\}, \\ \alpha_{res,n} + \sum_{j=int,rrt,rrt\&int} \alpha_{res,j} p_j, & \text{if } i = res, \\ \alpha_{f,res} p_{res} + \alpha_{f,res\&rrt} p_{res\&rrt}, & \text{if } i = f, \\ \alpha_{a,f} p_f + \alpha_{a,rrt\&f} p_{rrt\&f}, & \text{if } i = a. \end{cases} \tag{3.6}$$

Finally, since both direct and connecting routes could exist between response j and response i, define ρ_{ij} as the routing probability from j to i, which implies the probability

a patient going through these two services. Then, we obtain

$$\rho_{ij} = \begin{cases} 0, & \text{if there exists no path from service } j \text{ to service } i, \\ \alpha_{i,j}, & \text{if there exists a direct path from service } j \text{ to service } i, \\ \alpha_{i,k_m} \cdots \alpha_{k_1,j}, & \text{if there exist connecting paths from service } j \text{ to service } i, \end{cases} \tag{3.7}$$

where $k_1, \ldots, k_m$ denote the possible connecting services between j and i. Specifically, when there exists a direct path, we have

$$\begin{aligned} \rho_{ij} = \alpha_{ij}, \qquad & \text{if } i = res, j \in \{int, rrt, rrt\&int\}, \\ & \text{or } i = f, j = res \cup \{rrt\&res\}, \\ & \text{or } i = a, j = f \cup \{rrt\&f\}. \end{aligned}$$

When there exists a connecting path with one connecting service, we obtain

$$\begin{aligned} \rho_{ij} = \alpha_{ik_1}\alpha_{k_1 j}, \qquad & \text{if } i = a, k_1 = f, j = res \cup \{rrt\&res\} \\ & \text{or } i = f, k_1 = res, j \in \{int, rrt, rrt\&int\}. \end{aligned}$$

Finally, when the connecting path has two connecting services, it follows that

$$\rho_{ij} \quad = \quad \alpha_{ik_1}\alpha_{k_1,res}\alpha_{res,j}, i = a, k_1 = f, j \in \{int, rrt, rrt\&int\}.$$

Using the above notations, the performance of decision time (T_d and CV_d) can be evaluated:

Theorem 3.1 *Under assumptions (i)-(vii), the mean and coefficient of variation of*

decision time can be calculated as follows:

$$T_d = \sum_{i \in X} p_i \tau_i, \tag{3.8}$$

$$CV_d = \frac{1}{\sum_{i \in X} p_i \tau_i} \Bigg(\sum_{i \in X} p_i \tau_i^2 (cv_i^2 + 1 - p_i) - \sum_{i \in X} \tau_i \Big[\sum_{j \in X, j \neq i, n} (p_i - \rho_{ij}) p_j \tau_j \Big] \Bigg)^{\frac{1}{2}}, \tag{3.9}$$

where p_i and ρ_{ij} are defined in (3.6) and (3.7), respectively.

Proof: See Appendix A. ■

Theorem 3.1 provides a method to calculate the mean and CV of decision time, which enables us to evaluate the system performance under different scenarios. Note that the CV of decision time can be obtained without the knowledge of complete distributions of response times.

3.3.2 Response-time Performance

Exponential Case

As shown in Figure 3.1, a series of multiple responses could occur from declining to decision making. Then the response-time performance refers to the probability that the summation of a series of response times will be less than the desired care delivery time T_s. If each response time follows an exponential distribution, then the decision time which consists of a series of various exponential services will follow a hypoexponential distribution (Bolch et al. [15]). Thus, the probability that the decision is less than the desired care delivery time T_s can be calculated as follows:

Theorem 3.2 *Under assumptions (i)-(vii) and exponential response time for each service i, $i \in X$, the response-time performance, RTP, can be evaluated as follows:*

$$\begin{aligned} RTP &= \sum_{i\in X, j\neq f,a} \left(1 - e^{-\frac{T_s}{\tau_i}}\right)\alpha_{i,n} + \\ &\quad \sum_{i\in X} \sum_{j\in X, j\neq i,n} \left(1 - C_j e^{-\frac{T_s}{\tau_j}} - C_i e^{-\frac{T_s}{\tau_i}} - \sum_{l=1}^{m} C_{k_l} e^{-\frac{t}{\tau_{k_l}}}\right)\rho_{ij}\alpha_{j,n}, \end{aligned} \tag{3.10}$$

where $k_l \in X$, $l = 1, \ldots, m$, are the responses between services i and j, and m is the number of such responses. In addition,

$$C_i = \sum_{i\in X} \prod_{j\in X, j\neq i} \frac{\tau_j}{\tau_i - \tau_j}. \tag{3.11}$$

Proof: See Appendix A. ■

Remark 3.3 Unlike Theorem 3.1 which is applicable to any distribution, the results of Theorem 3.2 is only used for exponential response times. Extension of the study to non-exponential cases will be discussed next.

RTP Approximation

Theorem 3.2 provides a formula to calculate RTP in a rapid response system. However, the formula is very complicated and it requires detailed information of all the routing probabilities and response times. In practice, sometimes such information may not be always available. Thus, it may be more preferable to derive an approximation formula based on the total response time. In other words, consider all the possible response routes, if we aggregate all the responses weighted by routing probabilities, we obtain an aggregated response time, then an approximate formula to calculate RTP based on the aggregated time is needed.

Assume that the aggregated response time still follows exponential distribution, then RTP can be approximated by

$$RTP_{appr} = 1 - e^{-\frac{T_s}{T_d}}. \tag{3.12}$$

where T_d is calculated from Theorem 3.2.

Clearly, RTP_{appr} is only an approximation, since the aggregated response will not be exponential anymore. Therefore, numerical studies have been carried out to evaluate the accuracy of such approximation. A total of 100 systems have been investigated, each with a unique set of routing probabilities selected randomly. In each system, 100 sets of response time have been randomly selected for evaluation. Introduce Δ_{appr} to represent the relative difference between RTP and RTP_{appr}, i.e.,

$$\Delta_{appr} = \frac{|RTP - RTP_{appr}|}{RTP} \cdot 100\%.$$

In all the experiments, the difference between RTP calculated by Theorem 3.2 and by Equation (3.12), is small. The average of Δ_{appr} is only 0.6%, and the maximal one is 3%. An illustration of such approximation is shown in Figure 3.2. As shown in the figure, RTP_{appr} provides a close approximation to the actual response-time performance. Thus, (3.12) can provide an acceptable estimate of RTP, in particular, when detailed response information is not available.

Extension to Non-exponential Case

Theorem 3.2 assumes exponential response time for each care provider. In practice, such an assumption may not hold. Therefore, in this subsection, we extend the study to address non-exponential response time scenario. As an exact formula for response time with general distribution is impossible to derive, approximation is pursued here. The

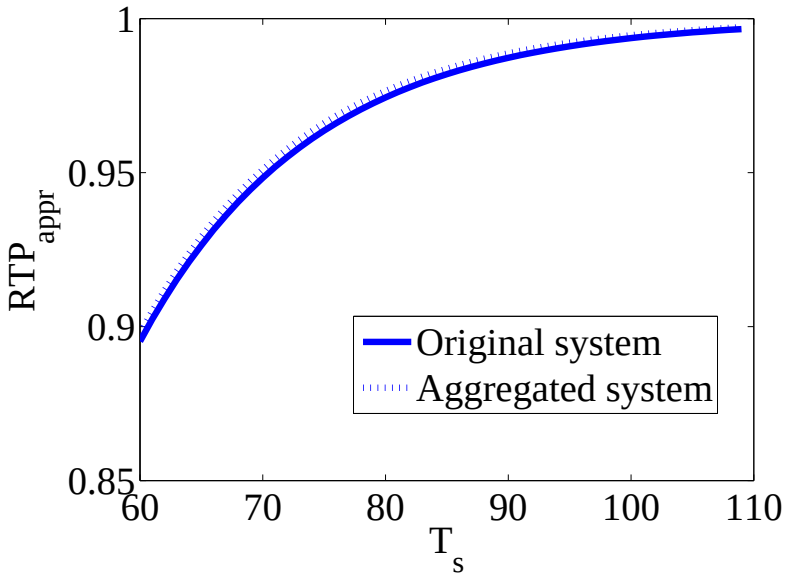

Figure 3.2: Approximation of RTP by aggregation

questions raised here are: first, will RTP be dependent or independent on the distribution types of response time? Second, if it is independent, does it exist an empirical formula to evaluate RTP as a function of the data that can be collected on the hospital floor?

Question 1: Will RTP be dependent or independent on the distribution types of response time?

To answer this question, we carry out numerical studies to verify the dependence or independence of RTP with respect to distribution type. Specifically, multiple networks with randomly selected routing probabilities are generated. In each network, responses with either lognormal or gamma distributions are assumed. The reason of selecting lognormal and gamma is that they both have two parameters which enable us to place the coefficient of variation freely. Specifically, the following five scenarios are considered:

- Lognormal: all responses follow lognormal distribution.
- Gamma: all responses follow gamma distribution.
- Mix 1: all responses involving RRT participation follow lognormal distribution, while others follow gamma distribution.
- Mix 2: all responses called by the primary nurse follow lognormal distribution, while others follow gamma distribution.
- Mix 3: the responses of intern, resident, fellow, attending, and RRT and attending follow lognormal distribution, while others follow gamma distribution.

In addition, the mean and CV are kept the same for each response under different scenarios. Then, for a given desired time period T_s, the RTPs are evaluated by using simulations in all scenarios, and then compared.

We first conduct ANOVA tests to investigate whether the RTP of the five systems are statistically indifferent(p>0.05). The results show that of p value is less than 0.05 in 2 out of 5 ANOVA tests, so we can not conclude that these systems are statistically indifferent. However, it should be noted that the simulation run time in each trial and the trial number affect the result of p value. Therefore we further investigate whether RTP is practically independent of the distribution type. Given a desire time T_s for each network, we compare RTPs of the five aforementioned scenarios, In each comparison, the largest relative difference δ is defined as:

$$\delta = \frac{\max_i RTP_i - \min_j RTP_j}{\min_j RTP_j} \cdot 100\%, i, j \in \{\text{Lognormal, Gamma, Mix 1, Mix 2, Mix 3}\},$$

where RTP_i represents the RTP in scenario i.

The comparison results are presented in Table 3.1. In all the comparisons, it has

Table 3.1: The comparison of RTPs among five networks, δ

T_s(minute)	70	60	50	40
Network 1	0.06%	0.10%	0.31%	0.55%
Network 2	0.26%	0.43%	0.88%	1.55%
Network 3	0.62%	0.91%	1.39%	1.95%
Network 4	0.31%	0.36%	0.67%	1.29%
Network 5	0.23%	0.37%	0.77%	1.43%

been observed that the difference, δ, is small. The average of δ is 0.72%, and the maximum one is 1.95%. An illustration of comparison among RTPs in one network among the above 5 scenarios is shown in Figure 3.3. These results indicate that the RTP is practically independent of the distribution type, but mainly depends on the mean and CV of response time.

Question 2: Does it exist an empirical formula to evaluate RTP as a function of the data that can be collected on the hospital floor?

Since each response may have different service time distribution or coefficient of variation, it is more convenient to group or aggregate those CVs so that we can use an universal CV for each response to discover an empirical formula to evaluate RTP in non-exponential response environment. Here we propose a simple approach to obtain this universal CV, referred to as a "weighted" CV. Specifically, for a rapid response system with i responses, the weighted CV can be expressed as follows:

$$cv_{weighted} = \frac{\sum_i p_i cv_i}{\sum_i p_i}, \tag{3.13}$$

where cv_i is the coefficient of variation of response i, and p_i is the probability that response i has been carried out, as defined earlier.

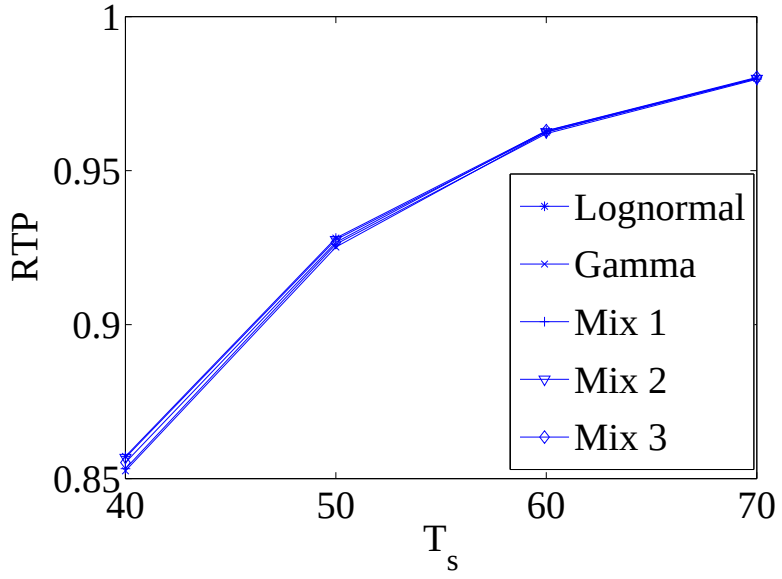

Figure 3.3: Impact of distribution type on RTP

To investigate the applicability of this weighted CV, we carry out numerical experiments to compare the RTPs in original system (denoted as $RTP_{original}$) and the RTPs obtained using such a weighted CV (denoted as $RTP_{weighted}$). Numerous examples are randomly generated for such experiments. To illustrate the results of these comparisons, the results from 5 networks are presented in Table 3.2. In each network, under the assumption of response distribution being lognormal and gamma, using 10 instances with different T_ss, the average relative difference of $RTP_{original,j}$ and $RTP_{weighted,j}$ is calculated, which is denoted as $\overline{\epsilon_i}$, we have

$$\overline{\epsilon_i} = \frac{\sum_{j=1}^{10} \frac{|RTP_{original,j} - RTP_{weighted,j}|}{RTP_{original,j}}}{10} \cdot 100\%, i \in \{\text{lognormal, gamma}\}.$$

As one can see, for lognormal distribution, the average $\epsilon_{lognormal}$ is 0.9%, while the largest one is 1.3%. For gamma assumption, such numbers become 1.2% and 1.7%,

Table 3.2: The difference of RTP by using a weighted CV, $\overline{\epsilon_i}$

System	1	2	3	4	5
$\overline{\epsilon_{lognormal}}$	1.1%	0.5%	0.4%	1.1%	1.3%
$\overline{\epsilon_{gamma}}$	1.7%	1.0%	0.6%	1.6%	1.1%

respectively. Similar results are observed in all other experiments, which shows that the weighted CV can be used for an empirical formula to evaluate RTP in non-exponential response scenarios.

Next, using the weighted CV, we derive an empirical formula to approximate the RTP in non-exponential case by proposing a linear interpolation approach . Specifically, we first calculate RTP_{exp} using formula (3.10) under exponential assumption. Then we calculate RTP_{fixed} for fixed response time, i.e., all response times are constants. This is carried out using the following formula:

$$RTP_{fixed} = \sum_{s_k \leq T_s} \gamma_k, \tag{3.14}$$

where s_k is the sum of all response time of route k, and γ_k is the probability of this route, which can be obtained by Equations (A.5) and (A.6) in the Proof of Theorem 3.2 in Appendix A.

For any other systems with weighted CV between 0 and 1, we estimate their RTP by a linear interpolation between RTP_{fixed} and RTP_{exp}, i.e.,

$$RTP_{est} = RTP_{fixed} - (RTP_{fixed} - RTP_{exp})cv_{weighted}. \tag{3.15}$$

In other words, for a rapid response system with an arbitrary distribution of response times, its RTP can be calculated using expressions (3.10), (3.14) and (3.15). An illustration of such estimation is shown in Figure 3.4.

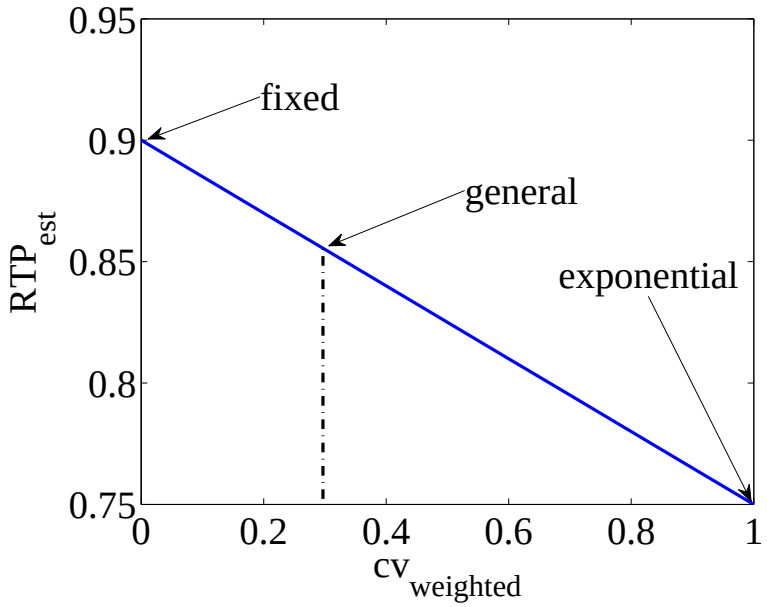

Figure 3.4: RTP approximation by linear interpolation

Similar to the numerical experiments introduced before, we investigate the accuracy of such estimate using linear interpolation through simulation experiments. Under lognormal and gamma assumptions of response time, respectively, the same five networks described above are used. In each network, with 10 different cvs between 0 and 1, including $cv = 0$ and $cv = 1$, the average relative difference under instance j between RTP obtained using simulation, $RTP_{sim,j}$, and RTP obtained using linear interpolation, $RTP_{cal,j}$, is calculated, which is denoted as $\overline{\delta_i}$ i.e.,

$$\overline{\delta_i} = \frac{\sum_{j=1}^{10} \frac{|RTP_{sim,j} - RTP_{cal,j}|}{RTP_{sim,j}}}{10} \cdot 100\%, i \in \{\text{lognormal, gamma}\}.$$

Table 3.3 shows the results.

Clearly, the approximation error is very small. For the five networks with lognormal and gamma distributions, the average values of $\delta_{lognormal}$ and δ_{gamma} are 0.7% and 0.8%,

Table 3.3: Accuracy of RTP estimation, $\overline{\delta_i}$

System	1	2	3	4	5
$\delta_{lognormal}$	1.1%	0.6%	0.5%	0.7%	0.9%
δ_{gamma}	1.9%	0.4%	0.3%	1.1%	0.9%

respectively. Similar results are obtained in all other experiments. Therefore, we conclude that the RTP estimation method introduced here provides a quantitative approach to estimate system response-time performance for any unimodal distribution of service times.

We also conduct paired t-test to investigate the justification of both the usage of weighed cv and accuracy of RTP estimations. The results of p values are not always greater than 0.05, so statistically the we can not make the same conclusion earlier. However, the design of simulation run time and trial run number have an impact on the conclusion. Therefore, from applicability perspectives, the comparison results show that the approximation approach can be used.

3.4 System Properties

Using the results obtained above, this section is devoted to investigating system-theoretic properties of T_d, CV_d, and RTP, i.e., analysis of the monotonicity of functions (3.1), (3.2), and (3.4) with respect to their arguments. The knowledge of the monotonicity properties will determine the variables, whose improvements lead to an decrease of T_d and CV_d, or an increase of RTP.

Proposition 3.1 *Under assumptions (i)-(vii), the expectation of decision time, T_d, is*

monotonically increasing with respect to τ_i, the mean of response time of care provider i.

Proof: See Appendix A. ■

Proposition 3.2 *Under assumptions (i)-(vii), the coefficient of variation of decision time, CV_d, is monotonically increasing with respect to cv_i, the coefficient of variation of response time of care provider i.*

Proof: See Appendix A. ■

Illustrations of such monotonicity properties are shown in Figures 3.5 and 3.6 for T_d and CV_d, respectively. As one can see, a linearly (or close to linearly) increasing behavior is observed for the mean and CV of decision time with respect to their arguments.

Remark 3.4 In addition to monotonicities of T_d and CV_d, one may concern about the impact of mean response time on the variability of decision time. It is easy to show that CV_d is independent to the changes of τ_i as long as cv_i remains unchanged. Specifically, by rewriting

$$\frac{\partial CV_d}{\partial \tau_i} = \frac{\partial CV_d}{\partial cv_i} \cdot \frac{\partial cv_i}{\partial \tau_i},$$

it follows that the second term equals to zero when cv_i is a constant. In other words, although reducing τ_i can lead to reduction of T_d, it will not result in a decrease of CV_d unless cv_i is reduced.

Similar property is observed for RTP as well, as shown below and in Figure 4.1, reducing each response time will lead to higher RTP.

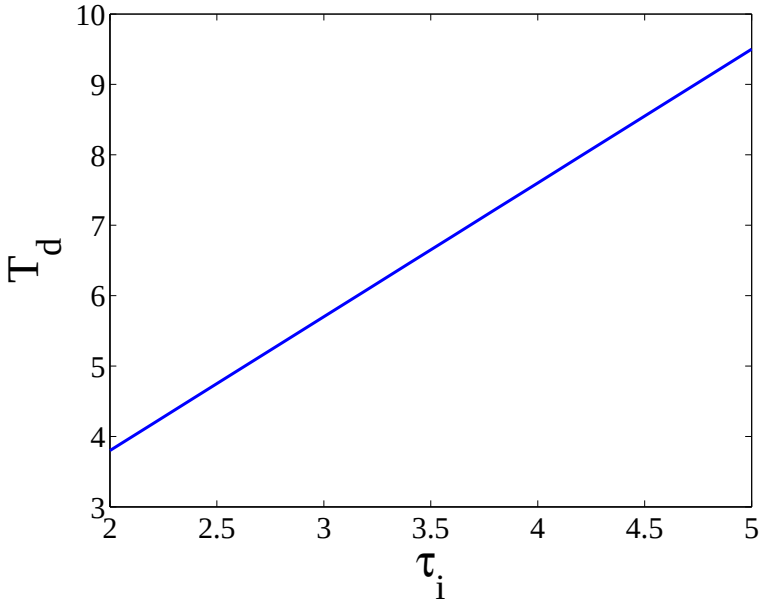

Figure 3.5: Monotonicity of T_d with respect to τ_i

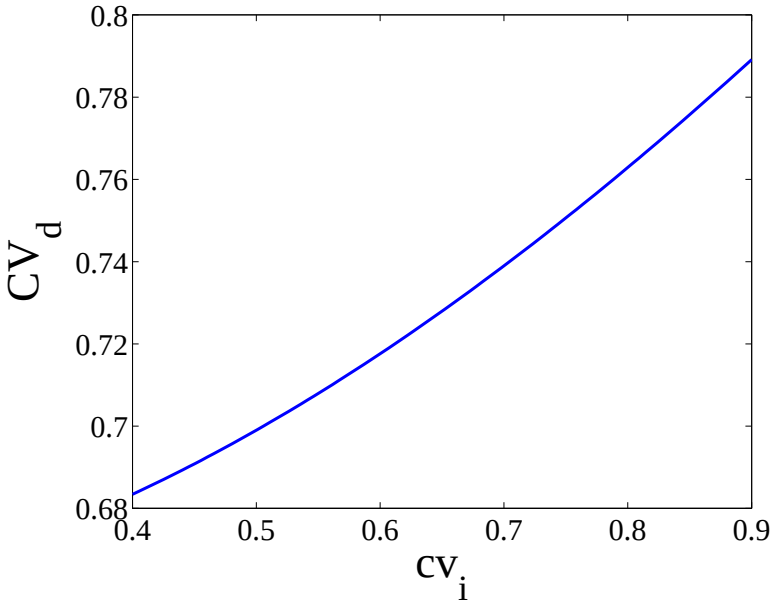

Figure 3.6: Monotonicity of CV_d with respect to cv_i

Proposition 3.3 *Under assumptions (i)-(vii), and exponential response time for each service i, RTP is monotonically decreasing with respect to τ_i, $i \in X$.*

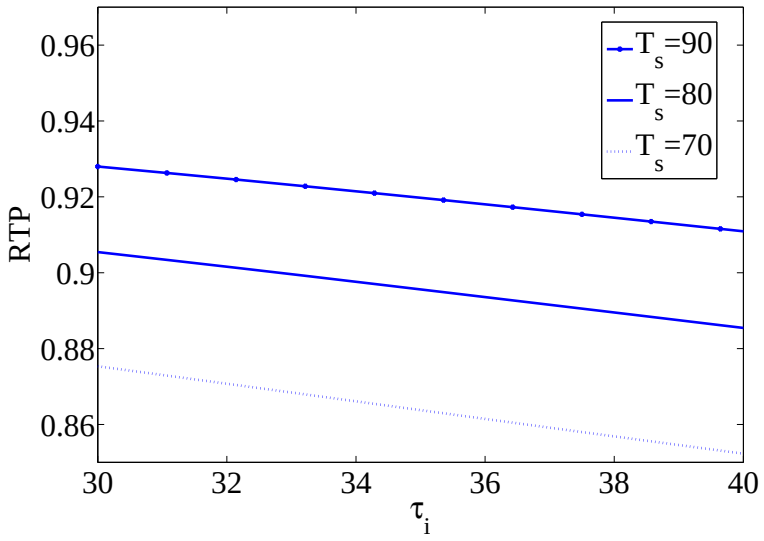

Figure 3.7: Monotonicity of RTP with respect to τ_i

The above monotonic properties match the intuitions well. This indicates that by reducing the response time and its variability, the decision time can be shortened and the variability of decision time can be reduced, respectively. Then, the next question is, which response or service should be focused on so that the maximum benefit (reduction of T_d or CV_d or increase of RTP) can be obtained. This leads to the bottleneck identification problem in continuous improvement, which will be addressed next.

3.5 Continuous Improvement

Bottleneck identification and mitigation have been viewed as the most effective way to improve system performance in manufacturing systems (Li and Meerkov [74]), where a

bottleneck machine is the one that impedes the system performance in the strongest manner. In healthcare delivery systems, a bottleneck process is referred to as the one whose improvement will lead to the largest improvement in performance measure. For example, it has been shown in Brenner et al. [18] and Zeng et al. [119] that diagnostic testing could be the bottleneck in emergency department from the patient flow perspective.

In this work, we study bottleneck responses to identify the most effective service to reduce decision time and its variability, or to increase response-time performance. In other words, the question is: which response time and its variability should be reduced so that it could lead to the largest reduction of decision time and its variability, or the largest increase in RTP? To answer these questions, we introduce the notions of bottleneck response for decision time (BN-τ) and its variability (BN-*cv*), and response-time performance bottleneck (BN-*rtp*).

3.5.1 Bottleneck Response for Decision Time and its Variability

The following is the definition of bottleneck responses.

Definition 3.1 *Under assumptions (i)-(vii), response i is the bottleneck response for decision time (BN-τ) if*

$$\frac{\partial T_d}{\partial \tau_i} > \frac{\partial T_d}{\partial \tau_j}, \qquad \forall i \neq j, \quad i, j \in X. \tag{3.16}$$

Definition 3.2 *Under assumptions (i)-(vii), response i is the bottleneck response for decision time variability (BN-cv) if*

$$\frac{\partial CV_d}{\partial cv_i} > \frac{\partial CV_d}{\partial cv_j}, \qquad \forall i \neq j, \quad i, j \in X. \tag{3.17}$$

That is, if the response time (or the CV of response time) of a process (e.g., RRT & Resident) is reduced, it can lead to the largest decrease in T_d (respectively, CV_d), then, this is the bottleneck response for decision time (respectively, for decision time variability).

The objective of continuous improvement is to develop methods for BN-τ and BN-*cv* identification, which will lead to structuring improvement projects directed to their mitigation. To identify a BN-τ, the following results are obtained:

Proposition 3.4 *Under assumptions (i)-(vii),*

$$\frac{\partial T_d}{\partial \tau_i} > \frac{\partial T_d}{\partial \tau_j}$$

if and only if

$$p_i > p_j, \quad \forall i \neq j, \quad i, j \in X.$$

Proof: See Appendix A. ■

Proposition 3.4 indicates that the bottleneck response for decision time, BN-τ, is the response with the largest p_i. Intuitively, a response with the largest p_i implies that it has the largest probability to be selected so that its weight is the highest. Therefore, a bottleneck indicator by measuring p_i is introduced.

BN-τ indicator: Define I_τ as the BN-τ indicator, where

$$I_{\tau,j} = p_j, \quad j \in X. \tag{3.18}$$

Then, the response that has the largest I_τ is the BN-τ.

Remark 3.5 As shown in Figure 3.1, the primary nurse is always the first provider responding to patient declining, thus, $p_n = 1$. This implies the nurse's response to

deteriorating signals is always the BN-τ, which indicates the importance of activation of response process. Then, the response with the second largest I_τ will be another bottleneck in the response process.

Next, the BN-cv is investigated. We obtain

Proposition 3.5 *Under assumptions (i)-(vii),*

$$\frac{\partial CV_d}{\partial cv_i} > \frac{\partial CV_d}{\partial cv_j}$$

if

$$p_i\tau_i^2 cv_i > p_j\tau_j^2 cv_j, \quad \forall i \neq j, \quad i, j \in X.$$

Proof: See Appendix A. ■

This result indicates that the bottleneck response variability, BN-cv, is the response that has the largest $p_i\tau_i^2 cv_i$. It can be understood as the response that has the largest product of entering probability, mean square and CV of response time. Reducing the variability of such a response will lead to the largest reduction in CV_d. Thus, the following indicator is introduced:

BN-cv indicator: Define I_{cv} as the BN-cv indicator, where

$$I_{cv,j} = p_j\tau_j^2 cv_j, \quad j \in X. \tag{3.19}$$

Then, the response that has the largest I_{cv} is the BN-cv.

Using these indicators, the BN-τ and BN-cv can be easily identified. Note that the calculation of such indicators can be carried out using the data collected on the hospital floor without calculations of performances and their partial derivatives.

3.5.2 Response-time Performance Bottleneck

Next, we investigate response-time performance bottlenecks. The RTP bottleneck (BN-*rtp*) is defined as

Definition 3.3 *Response i is the bottleneck response for RTP (BN-rtp) if*

$$\left|\frac{\partial RTP}{\partial \tau_i}\right| > \left|\frac{\partial RTP}{\partial \tau_j}\right|, \qquad \forall i \neq j, \quad i, j \in X. \tag{3.20}$$

Bottleneck Set Indicator

The objective of RTP continuous improvement is to develop a method for BN-*rtp* identification, which will lead to improving RTP in more effective manner. However, due to the complexity in RTP calculation, exact solution to discover BN-*rtp* is not available. Thus, an approximation approach is sought. Specifically, considering that the first levels of help called by the primary nurse have higher probability of activation than higher level physicians (such as fellow and attending doctors), we investigate a simplified structure that the decision is made only after one level of responses. In other words, assume

$$\alpha_{a,rrt\&f} = \alpha_{f,res} = \alpha_{f,rrt\&res} = \alpha_{res,int} = \alpha_{res,rrt} = \alpha_{res,rrt\&int} = 0. \tag{3.21}$$

Then, the following proposition can be derived:

Proposition 3.6 *Under assumptions (i)-(vi) and condition (3.21)*

$$\left|\frac{\partial RTP}{\partial \tau_i}\right| > \left|\frac{\partial RTP}{\partial \tau_j}\right|$$

if

$$e^{\frac{-T_s}{\tau_i}} \frac{\alpha_{i,n}}{\tau_i^2} > e^{\frac{-T_s}{\tau_j}} \frac{\alpha_{j,n}}{\tau_j^2}, \quad \forall i \neq j \in X, i, j \neq f, a.$$

Proof: See Appendix A. ■

Based on this result, we hypothesize that $e^{-T_s/\tau_j}\alpha_{j,n}/\tau_j^2$ can be used as a bottleneck set indicator to identify a set of BNs-*rtp* in the original system (i)-(vii). To verify such a hypothesis, numerical studies have been carried out. Specifically, for randomly selected routing probabilities $\alpha_{i,j}$s and desired care delivery time T_s, the service times τ_is are also randomly selected. From Definition 3.3, we obtain a list of $|\frac{\partial RTP}{\partial \tau_i}|$ ranking from high to low, which is referred to as the *true BN-rtp set*. Then by comparing $e^{-T_s/\tau_j}\alpha_{j,n}/\tau_j^2$, we obtain another set, referred to as the *hypothesized BN-rtp set*. Then a comparison between the sets can be carried out. Such experiments are repeated 1000 times.

From such a numerical study, we observe that in more than 90% of cases, the *true BN-rtp set* coincides with the *hypothesized BN-rtp set* for the top two BNs-*rtp*. In other words, the top two responses for RTP improvement can be identified by finding the two highest $e^{-T_s/\tau_j}\alpha_{j,n}/\tau_j^2$ with more than 90% successful rate. Therefore, we think $e^{-T_s/\tau_j}\alpha_{j,n}/\tau_j^2$ can be used as a practical indicator to identify the RTP bottleneck set, i.e.,

BN-rtp set indicator: Define I_{rtp} as the BN-*rtp* set indicator, where

$$I_{rtp,j} = \frac{\alpha_{j,n}}{\tau_j^2} e^{-\frac{T_s}{\tau_j}}, \quad j \in \{int, rrt, res, rrt\&int, rrt\&res, rrt\&f, rrt\&a\}. \quad (3.22)$$

Then, the responses having the largest and the second largest $I_{rtp,j}$ comprise the BN-*rtp* set.

After identifying the BN-*rtp* set, analysis need to be carried out to focus on these two responses to improve the efficacy of rapid response operations.

Bottleneck Indicator

The above set indicator enables us to identify the two candidates as BN-*rtp*. However, it does not specify which one could be the true bottleneck. To further investigate the effectiveness of this indicator, we define σ as the probability of calling for higher level help. In practice, such probability is not high. Therefore, we limit the discussion for small σ. Table 3.4 provides the accuracy of using $I_{rtp,j}$ as an indicator to identify the BN-*rtp* rather than the set. As one can see, when σ is less than 20%, the successful percentage to identify the true BN-*rtp* is 93%. The higher the σ, the lower the percentage. When σ is 50%, the perfentage falls just below 80%. However, even if the indicator does not identify the response with the largest partial derivative, the difference in improvement comparing with the response which is the true bottleneck is only 0.000016 to 0.00011 when σ ranges from 20% to 50%, respectively. Moreover, in most cases, such indicator identifies the response with the second largest partial derivative. As shown in Table 3.4, 93% of time the second one is identified when $\sigma = 20\%$. Only 0.7% of time it identifies a response which is ranked out of the top three. For example, even when $\sigma = 50\%$, among 22% of time it does not identify the correct bottleneck, but 80% of time the second ranked response is identified, and only 4% of time it indicates a response out of the top three.

Based on these, we believe that $e^{-T_s/\tau_j}\alpha_{j,n}/\tau_j^2$ can be used not only a BN-*rtp* set, but as a practical indicator of BN-*rtp* when the probability of calling help from higher level doctors is small, since in practice, such probability typically is not large. Thus, we have

BN-rtp indicator: Define I_{rtp} as the BN-*rtp* indicator. The response that has the largest $I_{rtp,j}$ is the BN-*rtp* when σ is small.

Table 3.4: Accuracy of BN-rtp indicator

Maximum further help probability σ	0.2	0.3	0.4	0.5
Identifying the true BN (%)	93%	85%	82%	78%
Improvement difference comparing with the true BN ($\times 10^{-4}$)	0.16	0.42	0.55	1.10
Identifying the 2nd BN (%)	93%	89%	82%	80%
Identifying the out-of-top-3 BN (%)	0.7%	0.8%	2%	4%

Remark 3.6 In non-exponential case, since a linear interpolation based on CV is used to approximate RTP, it does not affect the inequality in partial derivatives. Thus, the BN-rtp indictor introduced before still works.

3.6 Extension: Addressing Resource Sharing

In this section, we consider an extension under the current framework. Specifically, in the previous model, only one patient is allowed in the network. It should be noted that, in most of the practical settings, this assumption holds because it is not typical that more than one patients are experiencing deterioration simultaneously. However, there might be cases where more frequent declinings occur which raise a resource sharing issue. For example, during night shift, there will be limited number of clinicians and clinical staff. Thus, if new request is generated for a care provider or one of the two jointly activated providers who is currently attending a declining patient, the new patient has to wait until the current rescuing effort is finished. Consequently, the research question is how to estimate the new mean decision time when possible resource sharing results

in extra waiting time. Using the same analytical framework as a building block, we add the assumption that multiple patients may present in the network and the resource may be unavailable, as shown in Figure 3.8. Furthermore, Figure 3.9 provides an example. The overlapping bolded time period reflects that RRT is being shared during when two patients are in declining status at the same time. The goal is to develop a method to address this issue. Multiple factors contribute to the extra waiting time. First, a patient in hospital wards may decline at any time, resource sharing can only possibly occur when multiple patients are experiencing declining during the same time period. Second, even if multiple patients are declining, they may request different providers so that which resource is shared and how long the extra waiting time is are still not clear. This depends on the providers' probability to be called and their response time. Therefore, closed form formula to estimate waiting time is extremely difficult to obtain. Therefore, we develop a two-level shared resource iteration (SRI) method to evaluate the new mean decision time. To introduce such a method, an illustrative example is presented next.

3.6.1 An Illustrative Example

We consider an example with 3 patients. First, we define resource as a care provider or one of the two joint providers. Specifically, the resources include intern, RRT, resident, fellow and attending, belong to resource set X_1. Notice here X_1 does not include nurse in this section, that is, $X_1 = \{int, rrt, res, f, a\}$. As stated before, the factors contributing to the extra waiting time are categorized into two groups, simultaneous declining of patients and requesting for the same provider. Therefore, two iteration procedures are introduced. The level-1 iteration investigates the possibility that the same resource is requested by more than one declining patients at the same time and the extra waiting

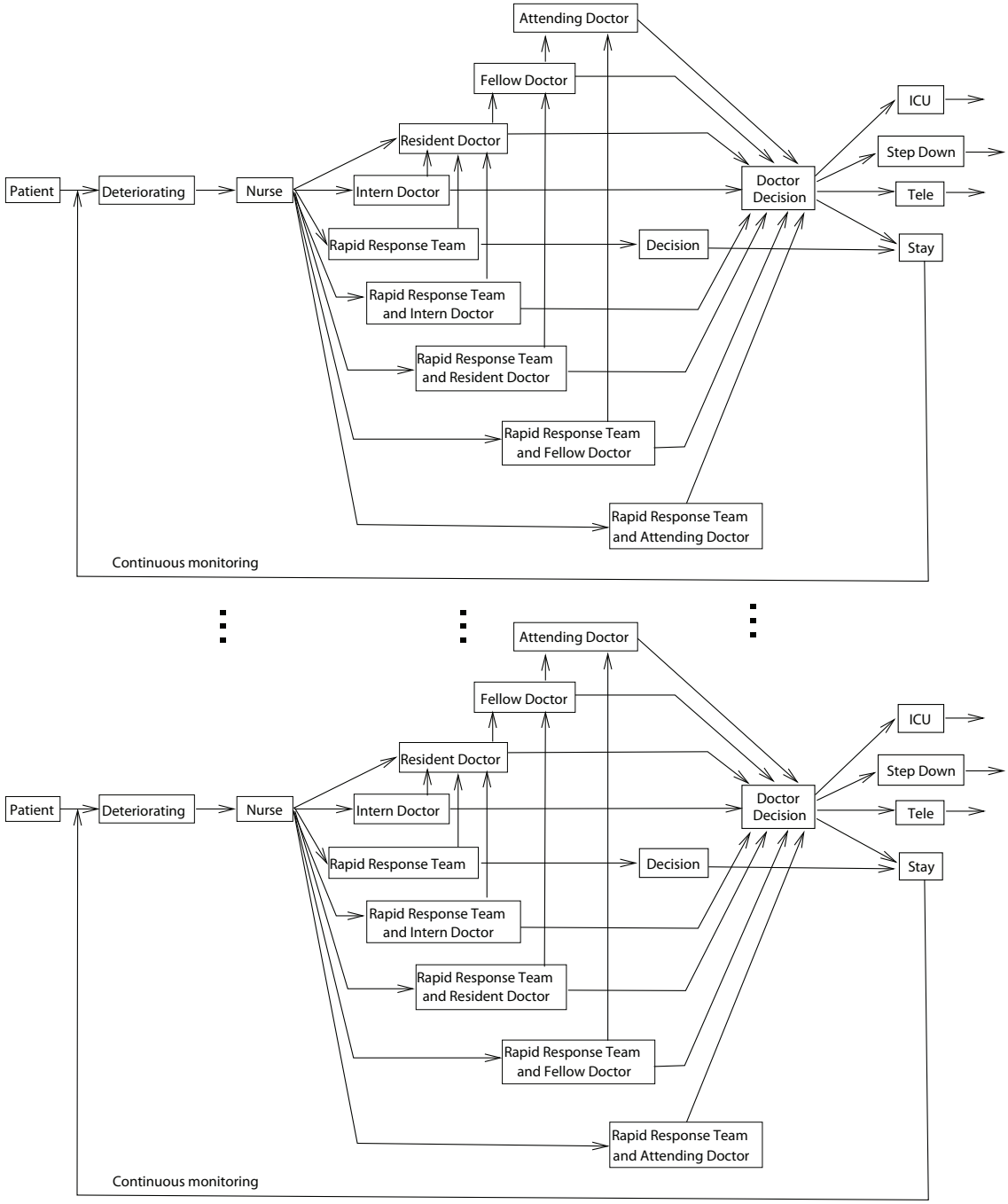

Figure 3.8: Multiple patients present in network

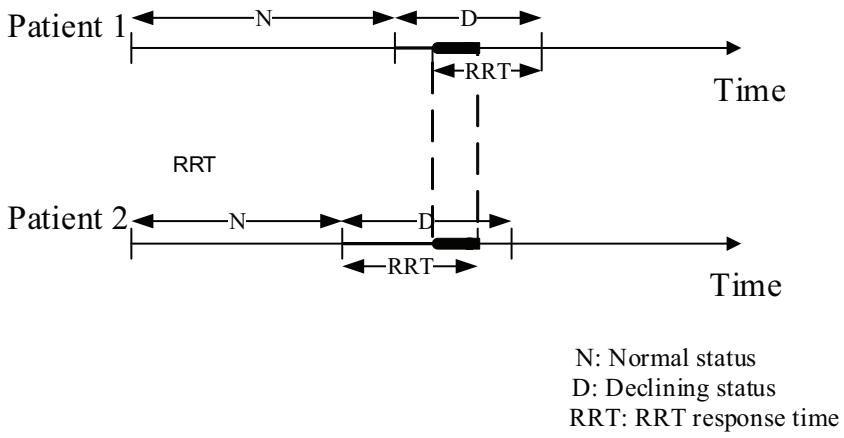

Figure 3.9: RRT is shared by two patients

time due to this. The level-2 iteration evaluates the probability that more than one patients are declining and its resulting waiting time. The results of level-1 iteration will be sent to level-2 iteration.

In the level-1 iteration, consider patient k, $k = 1, 2, 3$ and resource r, $r \in X_1$. Let $\tau_{k,r}^{(j)}$ denote the updated mean decision time including patient k's waiting time for resource r during the j-th iteration, $j = 1, 2, ...$ Introduce $p_{k,r}^{(j)}$ to denote the probability that patient k is occupying resource r when there is another request for the same resource during the j-th iteration. Lastly, from Equation (3.8) in Theorem 3.1, the mean decision time, under the assumption that all resources are available all the time, is denoted as T_d, while the probability and mean response time for response i, $i \in X = X_1 \cup X_2$, are p_i and τ_i, respectively.

At the beginning of the iteration, let all $\tau_{k,r}^{(0)}$ equal to T_d and $p_{k,r}^{(0)}$ equal to 0. For patient 1, the waiting will happen if he/she requests help from intern while patients 2 or 3 is being attended by the intern. This probability can be expressed as $p_{2,int}^{(0)} + p_{3,int}^{(0)}$. In addition, the average time that the intern provides rapid response can be expressed by $p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}$. This is due to the intern's role in both single provider (int)

and joint providers ($rrt\&int$) cases. Therefore, we update the mean decision time by including patient 1's waiting for intern, which is $\tau^{(1)}_{1,int}$, as follows.

$$\tau^{(1)}_{1,int} = T_d + (p^{(0)}_{2,int} + p^{(0)}_{3,int})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}). \tag{3.23}$$

We next update $p^{(1)}_{1,int}$, the probability that patient 1 is occupying the intern when there is another request for the resource. Again request for intern can happen in two cases: the single provider and joint providers. For the first case, $p_{int}\tau_{int}/\tau^{(1)}_{1,int}$ reflects the percentage of time the intern is occupied. Multiplying it by p_{int}, the probability the intern is working with another patient, we obtain the desired probability. Similarly, for the second case, we obtain $p^2_{rrt\&int}\tau_{rrt\&int}/\tau^{(1)}_{1,int}$. Therefore, the following formula is determined.

$$p^{(1)}_{1,int} = \frac{p^2_{int}\tau_{int} + p^2_{rrt\&int}\tau_{rrt\&int}}{\tau^{(1)}_{1,int}}. \tag{3.24}$$

Then, using $p^{(1)}_{1,int}$, the mean decision time contributed by patient 2's waiting for the intern, $\tau^{(1)}_{2,int}$, and the probability that patient 2 is occupying intern when the intern is requested by another patient, $p^{(1)}_{2,int}$, are updated.

$$\tau^{(1)}_{2,int} = T_d + (p^{(1)}_{1,int} + p^{(0)}_{3,int})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}), \tag{3.25}$$

$$p^{(1)}_{2,int} = \frac{p^2_{int}\tau_{int} + p^2_{rrt\&int}\tau_{rrt\&int}}{\tau^{(1)}_{2,int}}. \tag{3.26}$$

Lastly, using the above results, the update of patient 3 is completed similarly.

$$\tau^{(1)}_{3,int} = T_d + (p^{(1)}_{1,int} + p^{(1)}_{2,int})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}), \tag{3.27}$$

$$p_{3,int}^{(1)} = \frac{p_{int}^2 \tau_{int} + p_{rrt\&int}^2 \tau_{rrt\&int}}{\tau_{3,int}^{(1)}}. \tag{3.28}$$

This finishes the update related to the intern.

Next, we consider another resource, the resident doctor, with similar updating process from patient 1 to patient 3.

For patient 1:

$$\tau_{1,res}^{(1)} = T_d + (p_{2,res}^{(0)} + p_{3,res}^{(0)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}), \tag{3.29}$$

$$p_{1,res}^{(1)} = \frac{p_{res}^2 \tau_{res} + p_{rrt\&res}^2 \tau_{rrt\&res}}{\tau_{1,res}^{(1)}}. \tag{3.30}$$

For patient 2:

$$\tau_{2,res}^{(1)} = T_d + (p_{1,res}^{(1)} + p_{3,res}^{(0)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}), \tag{3.31}$$

$$p_{2,res}^{(1)} = \frac{p_{res}^2 \tau_{res} + p_{rrt\&res}^2 \tau_{rrt\&res}}{\tau_{2,res}^{(1)}}. \tag{3.32}$$

For patient 3, similar arguments apply. We have

$$\tau_{3,res}^{(1)} = T_d + (p_{1,res}^{(1)} + p_{2,res}^{(1)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}), \tag{3.33}$$

$$p_{3,res}^{(1)} = \frac{p_{res}^2 \tau_{res} + p_{rrt\&res}^2 \tau_{rrt\&res}}{\tau_{3,res}^{(1)}}. \tag{3.34}$$

We continue to update all the rest of resources as follows to complete the first iteration.

For RRT: from patients 1 to 3, we have

$$\begin{aligned}\tau^{(1)}_{1,rrt} &= T_d + (p^{(0)}_{2,rrt} + p^{(0)}_{3,rrt})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} \\ &\quad + p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a}), \qquad (3.35)\end{aligned}$$

$$\begin{aligned}p^{(1)}_{1,rrt} &= (p^2_{rrt}\tau_{rrt} + p^2_{rrt\&int}\tau^2_{rrt\&int} + p^2_{rrt\&res}\tau_{rrt\&res} \\ &\quad + p^2_{rrt\&f}\tau_{rrt\&f} + p^2_{rrt\&a}\tau_{rrt\&a})/\tau^{(1)}_{1,rrt}, \qquad (3.36)\end{aligned}$$

$$\begin{aligned}\tau^{(1)}_{2,rrt} &= T_d + (p^{(1)}_{1,rrt} + p^{(0)}_{2,rrt})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} \\ &\quad + p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a}), \qquad (3.37)\end{aligned}$$

$$\begin{aligned}p^{(1)}_{2,rrt} &= (p^2_{rrt}\tau_{rrt} + p^2_{rrt\&int}\tau^2_{rrt\&int} + p^2_{rrt\&res}\tau_{rrt\&res} \\ &\quad + p^2_{rrt\&f}\tau_{rrt\&f} + p^2_{rrt\&a}\tau_{rrt\&a})/\tau^{(1)}_{2,rrt}, \qquad (3.38)\end{aligned}$$

$$\begin{aligned}\tau^{(1)}_{3,rrt} &= T_d + (p^{(1)}_{1,rrt} + p^{(1)}_{2,rrt})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} \\ &\quad + p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a}), \qquad (3.39)\end{aligned}$$

$$\begin{aligned}p^{(1)}_{3,rrt} &= (p^2_{rrt}\tau_{rrt} + p^2_{rrt\&int}\tau^2_{rrt\&int} + p^2_{rrt\&res}\tau_{rrt\&res} \\ &\quad + + p^2_{rrt\&f}\tau_{rrt\&f} + p^2_{rrt\&a}\tau_{rrt\&a})/\tau^{(1)}_{3,rrt}. \qquad (3.40)\end{aligned}$$

For fellow, from patients 1 to 3, we have

$$\tau_{1,f}^{(1)} = T_d + (p_{2,f}^{(0)} + p_{3,f}^{(0)})(p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f}), \tag{3.41}$$

$$p_{1,f}^{(1)} = \frac{p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f}}{\tau_{1,f}^{(1)}}, \tag{3.42}$$

$$\tau_{2,f}^{(1)} = T_d + (p_{1,f}^{(1)} + p_{3,f}^{(0)})(p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f}), \tag{3.43}$$

$$p_{2,f}^{(1)} = \frac{p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f}}{\tau_{2,F}^{(1)}}, \tag{3.44}$$

$$\tau_{3,f}^{(1)} = T_d + (p_{1,f}^{(1)} + p_{2,f}^{(1)})(p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f}), \tag{3.45}$$

$$p_{3,f}^{(1)} = \frac{p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f}}{\tau_{3,f}^{(1)}}. \tag{3.46}$$

Finally, for attending, from patients 1 to 3, we have

$$\tau_{1,a}^{(1)} = T_d + (p_{2,a}^{(0)} + p_{3,a}^{(0)})(p_a\tau_a + p_{rrt\&a}\tau_{rrt\&a}), \tag{3.47}$$

$$p_{1,a}^{(1)} = \frac{p_a^2\tau_a + p_{rrt\&a}^2\tau_{rrt\&a}}{\tau_{1,A}^{(1)}}, \tag{3.48}$$

$$\tau_{2,a}^{(1)} = T_d + (p_{1,a}^{(1)} + p_{3,a}^{(0)})(p_a\tau_a + p_{rrt\&a}\tau_{rrt\&a}), \tag{3.49}$$

$$p_{2,a}^{(1)} = \frac{p_a^2\tau_a + p_{rrt\&a}^2\tau_{rrt\&a}}{\tau_{2,a}^{(1)}}, \tag{3.50}$$

$$\tau_{3,a}^{(1)} = T_d + (p_{1,a}^{(1)} + p_{2,a}^{(1)})(p_a\tau_a + p_{rrt\&a}\tau_{rrt\&a}), \tag{3.51}$$

$$p_{3,a}^{(1)} = \frac{p_a^2 \tau_a + p_{rrt\&a}^2 \tau_{rrt\&a}}{\tau_{3,a}^{(1)}}. \tag{3.52}$$

After finishing the first iteration, the above updated variables are used for subsequent iterations and the process is repeated if converges. More specifically, the following criteria is used for convergence:

Let $\delta = 10^{-5}$, if

$$|\tau_{i,r}^{(j+1)} - \tau_{i,r}^{(j)}| \leq \delta, i = 1, 2, 3, \tag{3.53}$$

$$|p_{i,r}^{(j+1)} - p_{i,r}^{(j)}| \leq \delta, i = 1, 2, 3. \tag{3.54}$$

Then we claim the procedure is convergent. When the procedure is convergent, we denote

$$\lim_{j \to \infty} \tau_{i,r}^{(j)} = \tau_{i,r}, i = 1, 2, 3, \tag{3.55}$$

$$\lim_{j \to \infty} p_{i,r}^{(j)} = p_{i,r}, i = 1, 2, 3. \tag{3.56}$$

Moreover, all $\tau_{i,r}, i = 1, 2, 3$, should be identical and all $p_{i,r}, i = 1, 2, 3$, will be the same, due to identical parameters in each patients. Therefore, upon convergence, we have

$$\tau_{1,r} = \tau_{2,r} = \tau_{3,r} = \mathrm{T}_r, \tag{3.57}$$

$$p_{1,r} = p_{2,r} = p_{3,r} = P_r. \tag{3.58}$$

After convergence, we obtain the output of the level-1 iteration T_{in}, which is the updated mean decision time by including the extra waiting time for all resources, due to sharing.

$$T_{in} = T_d + \Sigma_{r,r\in X_1} P_r \mathrm{T}_r. \tag{3.59}$$

Using the T_{in} from level-1 iteration, we start the level-2 iteration. Let $g_k^{(l)}$, $k = 1, 2, 3$, $l = 1, 2, ...$, denote the time proportion the patient is in declining status in iteration j, and $\mu_k^{(l)}$, $k = 1, 2, 3$, $l = 1, 2, ...$, characterize the newly updated mean decision time in iteration j after considering the time proportion that patient k is in declining status. Introduce T_{normal} to denote the average time the patient is in normal status. To start the iteration, assume all $g_k^{(0)}$ equal to 0 and $\mu_k^{(0)}$ to T_{in}. In the first iteration, for patient 1, the waiting will only occur if patient 1 is declining, while patient 2 or patient 3 is also declining, which can be expressed as $g_1^{(0)}(g_2^{(0)} + g_3^{(0)})$. Therefore we first obtain the updated $\mu_1^{(1)}$ as follows.

$$\mu_1^{(1)} = T_{in}(1 + g_1^{(0)}(g_2^{(0)} + g_3^{(0)})). \tag{3.60}$$

In addition, the time proportion patient 1 is in declining status is updated as

$$g_1^{(1)} = \frac{\mu_1^{(1)}}{\mu_1^{(1)} + T_{normal}}. \tag{3.61}$$

Moving on to patient 2 and patient 3 using the same logic, we obtained the new $\mu_2^{(1)}$, $\mu_3^{(1)}$, $g_2^{(1)}$ and $g_3^{(1)}$ below. For patient 2:

$$\mu_2^{(1)} = T_{in}(1 + g_2^{(0)}(g_1^{(1)} + g_3^{(0)})), \tag{3.62}$$

$$g_2^{(1)} = \frac{\mu_2^{(1)}}{\mu_2^{(1)} + T_{normal}}. \tag{3.63}$$

For patient 3:

$$\mu_3^{(1)} = T_{in}(1 + g_3^{(0)}(g_1^{(1)} + g_2^{(1)})), \tag{3.64}$$

$$g_3^{(1)} = \frac{\mu_3^{(1)}}{\mu_3^{(1)} + T_{normal}}. \tag{3.65}$$

This finishes the first iteration. Using the results of $g_i^{(1)}$ and $\mu_i^{(1)}$ from this iteration, we continue this process of updating by $g_i^{(l)}$ and $\mu_i^{(l)}$ considering patients 1, 2, 3 again, and repeat the process until the procedure converges. The convergence criteria is met when the following conditions hold:

$$|\mu_i^{(j+1)} - \mu_i^{(l)}| \leq \delta, i = 1, 2, 3, \tag{3.66}$$

$$|g_i^{(j+1)} - g_i^{(l)}| \leq \delta, i = 1, 2, 3. \tag{3.67}$$

where $\delta = 10^{-5}$. Again, if the procedure converges, we obtain

$$\lim_{l \to \infty} \mu_i^{(l)} = \mu_i, i = 1, 2, 3, \tag{3.68}$$

$$\lim_{l \to \infty} g_i^{(l)} = g_i, i = 1, 2, 3. \tag{3.69}$$

Moreover, all $\mu_i, i = 1, 2, 3$, should be the same, due to identical parameters in each patients. Therefore, let T_{final} denote the mean decision time, which is the final result, we have

$$\mu_1 = \mu_2 = \mu_3 = T_{final}. \tag{3.70}$$

After convergence of the above level-2 iteration, all issues related to resource sharing, which contribute to the addition of extra waiting time are addressed. An illustration of such a procedure is shown in Figure 3.10 where n patients present in the network. Numerical experiments have indicated that such procedure always converge.

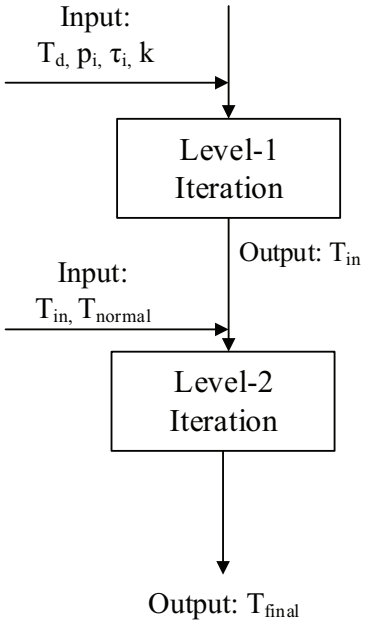

Figure 3.10: Illustration of two-level shared iteration procedure

3.6.2 General Procedure

Similar to the idea introduced in the illustrative example, we present here the general procedure. Using the introduced variables in the 3-patient illustrative example, the iteration procedure can be formally described as follows:

Procedure 3.2 ***Two-level Shared Resource Iteration***

(1) Level-1 iteration

Step 1.1: Initialization:

From Equation (3.6) and Equation (3.8), obtain p_i, $i \in X = X_1 \cup X_2$, *and* T_d. *Set* $j = 0$. *Set* $\tau_{k,r}^{(j)}$ *and* $p_{k,r}^{(j)}$ *equal to 0.*

Step 1.2: Update $\tau_{k,r}^{(j)}$ and $p_{k,r}^{(j)}$:

For patient 1,

For $r = int$,

$$\tau_{1,r}^{(j+1)} = T_d + \Sigma_{i=2,\ldots,n} p_{i,r}^{(j)} (p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}), \tag{3.71}$$

$$p_{1,r}^{(j+1)} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}}{\tau_{1,r}^{(j+1)}}. \tag{3.72}$$

For $r = res$,

$$\tau_{1,r}^{(j+1)} = T_d + \Sigma_{i=2,\ldots,n} p_{i,r}^{(j)} (p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}), \tag{3.73}$$

$$p_{1,r}^{(j+1)} = \frac{p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}}{\tau_{1,r}^{(j+1)}}. \tag{3.74}$$

For $r = rrt$,

$$\tau_{1,r}^{(j+1)} = T_d + \Sigma_{i=2,\ldots,n} p_{i,r}^{(j)} (p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a}), \tag{3.75}$$

$$p_{1,r}^{(j+1)} = (p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res} + p_{rrt\&f}^2\tau_{rrt\&f} + p_{rrt\&a}^2\tau_{rrt\&a})/\tau_{1,r}^{(j+1)}. \tag{3.76}$$

For $r = f$,

$$\tau_{1,r}^{(j+1)} = T_d + \Sigma_{i=2,\ldots,n} p_{i,r}^{(j)} (p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f}), \tag{3.77}$$

$$p_{1,r}^{(j+1)} = \frac{p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f}}{\tau_{1,r}^{(j+1)}}. \tag{3.78}$$

For $r = a$,

$$\tau_{1,r}^{(j+1)} = T_d + \Sigma_{i=2,\ldots,n} p_{i,r}^{(j)}(p_a\tau_a + p_{rrt\&a}\tau_{rrt\&a}), \tag{3.79}$$

$$p_{1,r}^{(j+1)} = \frac{p_a^2\tau_a + p_{rrt\&a}^2\tau_{rrt\&a}}{\tau_{1,r}^{(j+1)}}. \tag{3.80}$$

For patient $k = 2, .., n-1$,

For $r = int$,

$$\tau_{k,r}^{(j+1)} = T_d + (\Sigma_{i=1,\ldots,k-1} p_{i,r}^{(j+1)} + \Sigma_{i=k+1,\ldots,n} p_{i,r}^{(j)})(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}), \tag{3.81}$$

$$p_{k,r}^{(j+1)} = \frac{p_{int}^2\tau_{int} + p_{rrt\&int}^2.\tau_{rrt\&int}}{\tau_{k,r}^{(j+1)}}. \tag{3.82}$$

For $r = res$,

$$\tau_{k,r}^{(j+1)} = T_d + (\Sigma_{i=1,\ldots,k-1} p_{i,r}^{(j+1)} + \Sigma_{i=k+1,\ldots,n} p_{i,r}^{(j)})(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}), \tag{3.83}$$

$$p_{k,r}^{(j+1)} = \frac{p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res}}{\tau_{k,r}^{(j+1)}}. \tag{3.84}$$

For $r = rrt$,

$$\begin{aligned}\tau_{k,r}^{(j+1)} = T_d + (\Sigma_{i=1,\ldots,k-1} p_{i,r}^{(j+1)} + \Sigma_{i=k+1,\ldots,n} p_{i,r}^{(j)})(p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} \\ + p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a}),\end{aligned} \tag{3.85}$$

$$\begin{aligned}p_{k,r}^{(j+1)} = (p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res} \\ + p_{rrt\&f}^2\tau_{rrt\&f} + p_{rrt\&a}^2\tau_{rrt\&a})/\tau_{k,r}^{(j+1)}.\end{aligned} \tag{3.86}$$

For $r = f$,

$$\tau_{k,r}^{(j+1)} = T_d + (\Sigma_{i=1,\ldots,k-1} p_{i,r}^{(j+1)} + \Sigma_{i=k+1,\ldots,n} p_{i,r}^{(j)})(p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f}), \tag{3.87}$$

$$p_{k,r}^{(j+1)} = \frac{p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f}}{\tau_{k,r}^{(j+1)}}. \tag{3.88}$$

For $r = a$,

$$\tau_{k,r}^{(j+1)} = T_d + (\Sigma_{i=1,\ldots,k-1} p_{i,r}^{(j+1)} + \Sigma_{i=k+1,\ldots,n} p_{i,r}^{(j)})(p_a \tau_a + p_{rrt\&a}\tau_{rrt\&a}), \tag{3.89}$$

$$p_{k,r}^{(j+1)} = \frac{p_a^2 \tau_a + p_{rrt\&a}^2 \tau_{rrt\&a}}{\tau_{k,r}^{(j+1)}}. \tag{3.90}$$

For patient $k = n$,

For $r = int$,

$$\tau_{k,r}^{(j+1)} = T_d + \Sigma_{i=1,\ldots,n-1} p_{i,r}^{(j+1)}(p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}), \tag{3.91}$$

$$p_{k,r}^{(j+1)} = \frac{p_{int}^2 \tau_{int} + p_{rrt\&int}^2 . \tau_{rrt\&int}}{\tau_{k,r}^{(j+1)}} \tag{3.92}$$

For $r = res$,

$$\tau_{k,r}^{(j+1)} = T_d + \Sigma_{i=1,\ldots,n-1} p_{i,r}^{(j+1)}(p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res}), \tag{3.93}$$

$$p_{k,r}^{(j+1)} = \frac{p_{res}^2 \tau_{res} + p_{rrt\&res}^2 \tau_{rrt\&res}}{\tau_{k,r}^{(j+1)}}. \tag{3.94}$$

For $r = rrt$,

$$\begin{aligned}\tau_{k,r}^{(j+1)} = T_d + \Sigma_{i=1,\ldots,n-1} p_{i,r}^{(j+1)}(&p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} \\ &+ p_{rrt\&res}\tau_{rrt\&res} + p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a}),\end{aligned} \tag{3.95}$$

$$\begin{aligned}p_{k,r}^{(j+1)} = (&p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res} \\ &+ p_{rrt\&f}^2\tau_{rrt\&f} + p_{rrt\&a}^2\tau_{rrt\&a})/\tau_{k,r}^{(j+1)}.\end{aligned} \tag{3.96}$$

For $r = f$,

$$\tau_{k,r}^{(j+1)} = T_d + \Sigma_{i=1,\ldots,n-1} p_{i,r}^{(j+1)}(p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f}), \tag{3.97}$$

$$p_{k,r}^{(j+1)} = \frac{p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f}}{\tau_{k,r}^{(j+1)}}. \tag{3.98}$$

For $r = a$,

$$\tau_{k,r}^{(j+1)} = T_d + \Sigma_{i=1,...,n-1} p_{i,r}^{(j+1)} (p_a \tau_a + p_{rrt\&a} \tau_{rrt\&a}), \tag{3.99}$$

$$p_{k,r}^{(j+1)} = \frac{p_a^2 \tau_a + p_{rrt\&a}^2 \tau_{rrt\&a}}{\tau_{k,r}^{(j+1)}}. \tag{3.100}$$

Step 1.3: Iteration:

Let $j = j + 1$. Go back to Step 1.2 until all the stopping criteria is met. For a given $\delta = 10^{-5}$, we terminate the level-1 iteration until:

$$|\tau_{i,r}^{(j+1)} - \tau_{i,r}^{(j)}| \leq \delta, i = 1, 2, ..., n, \tag{3.101}$$

$$|p_{i,r}^{(j+1)} - p_{i,r}^{(j)}| \leq \delta, i = 1, 2, ..., n. \tag{3.102}$$

Step 1.4: Termination:

When the above terminating conditions are met, let

$$\tau_{1,r} = \tau_{2,r} = ... = \tau_{n,r} = \mathrm{T}_r, \tag{3.103}$$

$$p_{1,r} = p_{2,r} = ... = p_{n,r} = P_r. \tag{3.104}$$

$$T_{in} = T_d + \Sigma_{r,r \in X_1} P_r \mathrm{T}_r \tag{3.105}$$

(2) Level-2 iteration

Step 2.1: Initialization:

Set $l = 0$. Set $g_1^{(l)} = 0$ and $\mu_1^{(l)} = T_{in}$.

Step 2.2: Update $g_k^{(l)}$ and $\mu_k^{(l)}$:

For patient 1,

$$\mu_1^{(l+1)} = T_{in}(1 + g_1^{(l)} \Sigma_{i=2,...,n} g_i^{(l)}), \tag{3.106}$$

$$g_1^{(l+1)} = \frac{\mu_1^{(l+1)}}{\mu_1^{(l+1)} + T_{normal}}. \tag{3.107}$$

For patient $k = 2, ..., n-1$,

$$\mu_k^{(l+1)} = T_{in}(1 + g_k^{(l)}(\Sigma_{i=1,...,k-1} g_i^{(l+1)} + \Sigma_{i=k+1,...,n} g_i^{(l)})), \tag{3.108}$$

$$g_k^{(l+1)} = \frac{\mu_k^{(l+1)}}{\mu_k^{(l+1)} + T_{normal}}. \tag{3.109}$$

For patient $k = n$,

$$\mu_k^{(l+1)} = T_{in}(1 + g_k^{(l)} \Sigma_{i=1,...,n-1} g_i^{(l+1)}), \tag{3.110}$$

$$g_k^{(l+1)} = \frac{\mu_k^{(l+1)}}{\mu_k^{(l+1)} + T_{normal}}. \tag{3.111}$$

Step 2.3: Iteration:

Let $l = l + 1$. *Go back to Step 2.2 until all the stopping criteria is met.*

$$|\mu_i^{(l+1)} - \mu_i^{(l)}| \leq \delta, i = 1, 2, ..., n, \tag{3.112}$$

$$|g_i^{(l+1)} - g_i^{(l)}| \leq \delta, i = 1, 2, ..., n. \tag{3.113}$$

Step 2.4: Termination:

When the stopping condition is met, we have

$$\mu_i^{(l+1)} = \mu_i, i = 1, 2, ..., n. \tag{3.114}$$

Then,

$$\mu_1 = \mu_2 = ... = \mu_n = T_{final}. \tag{3.115}$$

Such a procedure includes two algorithms: level-1 iteration and level-2 iteration. The level-1 iteration can be mathematically proved to be convergent if number of patients in the network equals to 2. The convergence of level-2 iteration can be rigorously proved without the limitation of patient number. Theorems 3.3 and 3.4 are presented below.

Theorem 3.3 *When $n = 2$, the level-1 iteration of Procedure 3.2 is convergent. The following limits exist:*

$$\lim_{j\to\infty} \tau_{i,r}^{(j)} = \tau_{i,r}, i = 1, 2, \tag{3.116}$$

$$\lim_{j\to\infty} p_{i,r}^{(j)} = p_{i,r}.i = 1, 2. \tag{3.117}$$

Theorem 3.4 *The level-2 iteration of Procedure 3.2 is convergent, i.e., the following limits exist:*

$$\lim_{l\to\infty} \mu_i^{(l)} = \mu_i, i = 1, 2, ...n, \tag{3.118}$$

$$\lim_{l\to\infty} g_i^{(l)} = g_i, i = 1, 2, ...n. \tag{3.119}$$

Proof: See Appendix A. ■

If more than two patients present in the network, it is extremely difficult to provide a mathematical proof for level-1 iteration due to its nature of oscillating pattern. Numerical investigation of the convergence of such a procedure is conducted. Numerous examples are generated by randomly selecting parameters. In all the examples, this iterative approach is convergent and results in a unique solution. Therefore, we formulate it as a numerical fact:

Numerical Fact 3.1 *: The level-1 iteration of Procedure 3.2 is convergent when more than two patients present in the network, i.e., we have*

$$\lim_{j\to\infty} \tau_{i,r}^{(j)} = \tau_{i,r}, i = 1, 2, ..., n, \tag{3.120}$$

$$\lim_{j\to\infty} p_{i,r}^{(j)} = p_{i,r}.i = 1, 2, ..., n. \tag{3.121}$$

The following figures illustrate the convergence in the numerical fact, using a 3-patient example. Figures 3.11 to 3.20 show the convergence of $\tau_{i,r}$ and $p_{i,r}$ in the level-1 iteration of all resources. Figures 3.21 to 3.22 show the convergence of μ_i and g_i in the level-2 iteration. As can be seen in the figures, all the variables rapidly converge in less than 6 iterations. Moreover, the oscillating range is extremely small after the first iteration. For $\tau_{i,r}$s, at the first iteration, it is always true that $\tau_{3,r} > \tau_{2,r} > \tau_{1,r}$, for $r \in X_1$, however, starting from the second iteration, due to oscillation, this relationship no longer holds. After the third iteration, the differences between any two of the $\tau_{i,r}$s are too small to identify from the figures. For $p_{i,r}$s, at the first iteration, it is always true that $p_{3,r} < p_{2,r} < p_{1,r} >$ for $r \in X_1$, Similar to $\tau_{i,r}$s, starting from the second iteration, oscillation pattern makes this relationship no longer hold. Also similarly, the differences between any two of the $p_{i,r}$s are extremely small to tell from the figures after the third iteration. For μ_is and g_is, monotonicity property can be verified in the figures. Moreover, the difference between any of the two μ_is is always extremely small starting from the first iteration. Such small difference can be also observed regarding g_is.

3.6.3 Accuracy

The accuracy of the two-level shared resource iteration approach is examined by comparing the simulation results with those obtained by the iteration approach. In the

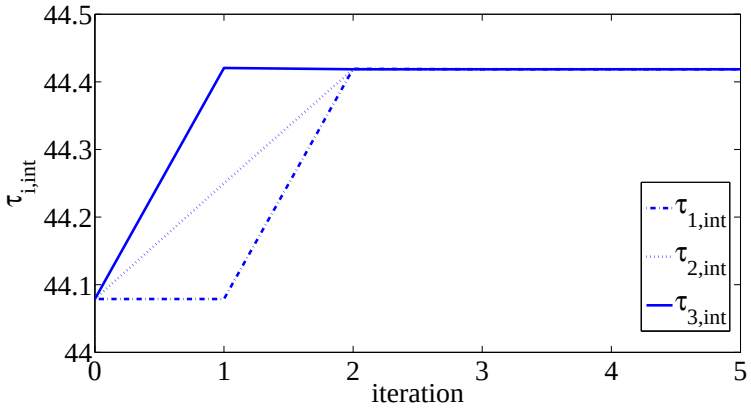

Figure 3.11: Convergence of $\tau_{i,int}$

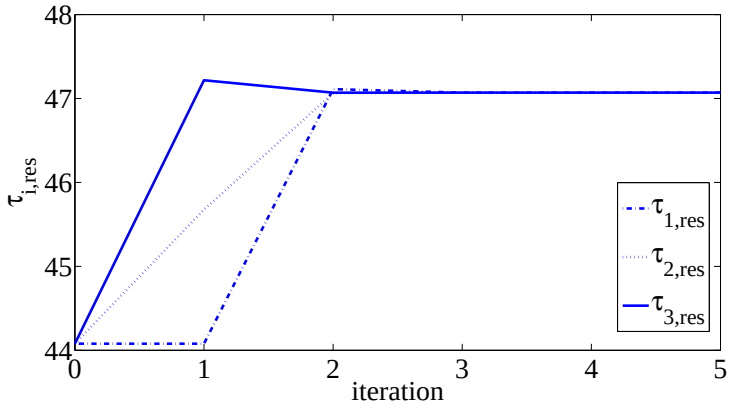

Figure 3.12: Convergence of $\tau_{i,res}$

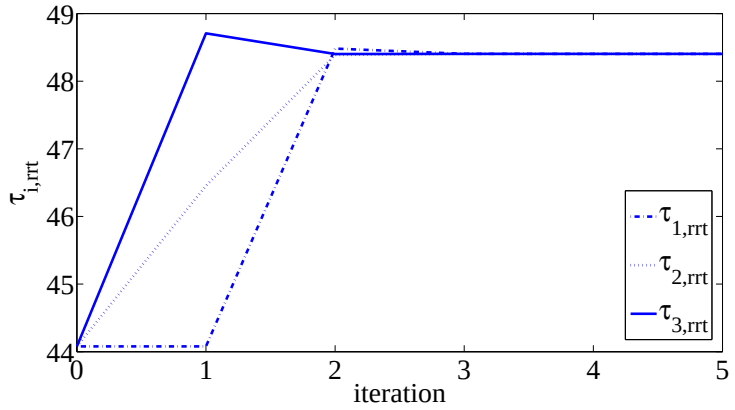

Figure 3.13: Convergence of $\tau_{i,rrt}$

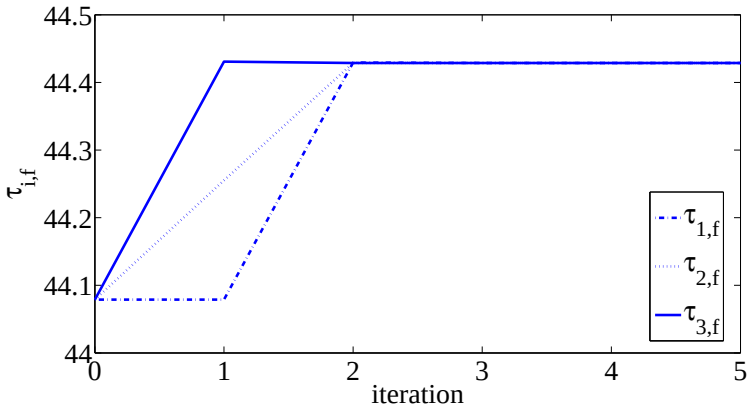

Figure 3.14: Convergence of $\tau_{i,f}$

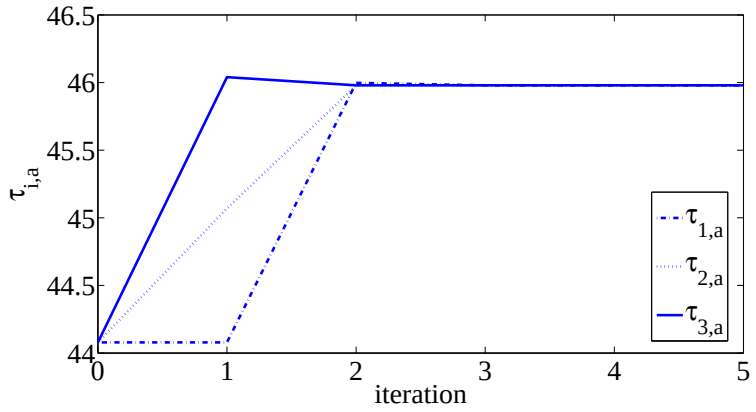

Figure 3.15: Convergence of $\tau_{i,a}$

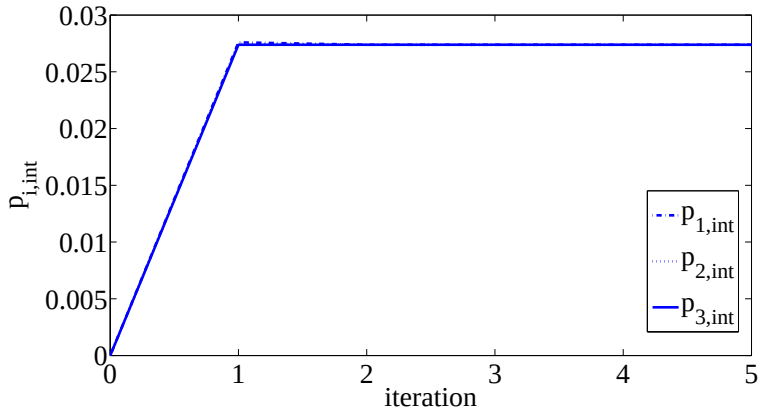

Figure 3.16: Convergence of $p_{i,int}$

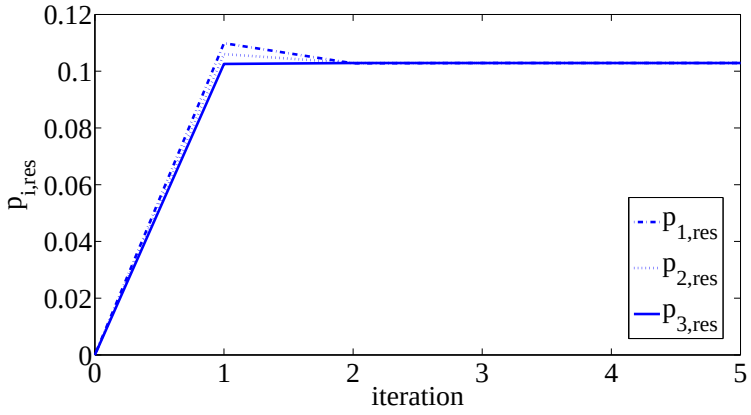

Figure 3.17: Convergence of $p_{i,res}$

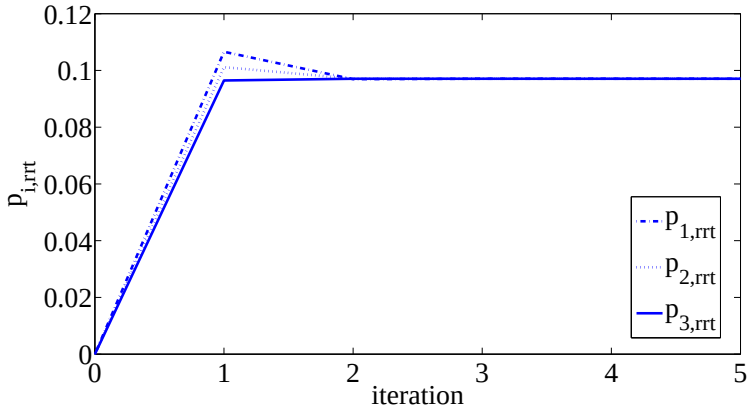

Figure 3.18: Convergence of $p_{i,rrt}$

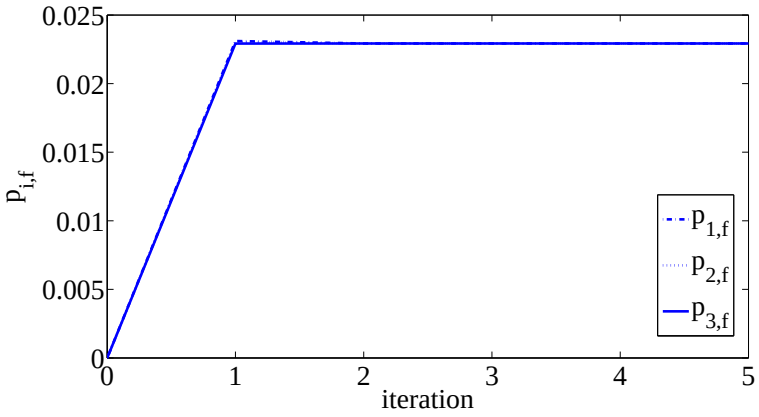

Figure 3.19: Convergence of $p_{i,f}$

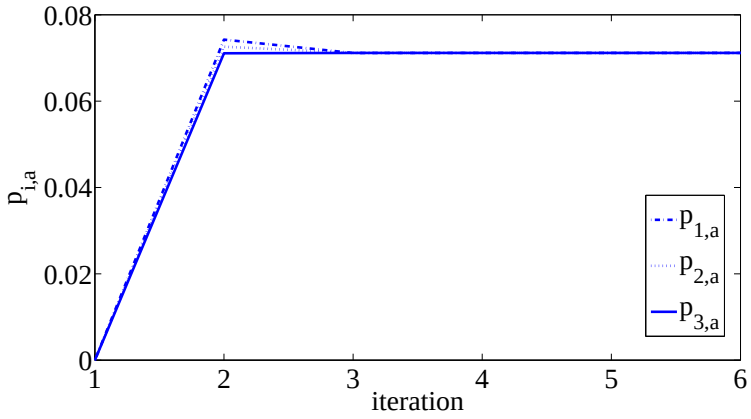

Figure 3.20: Convergence of $p_{i,a}$

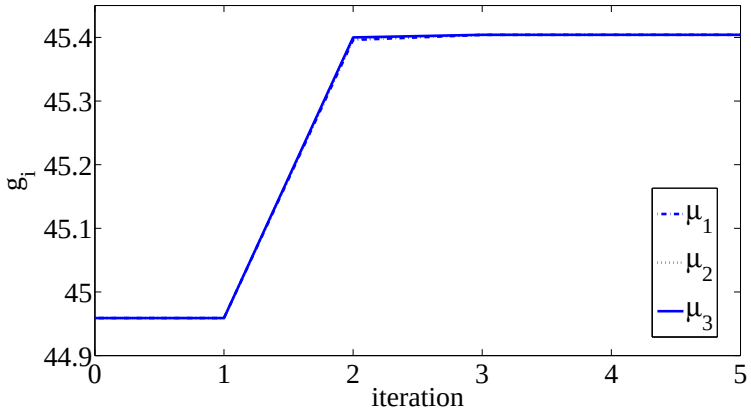

Figure 3.21: Convergence of μ_i

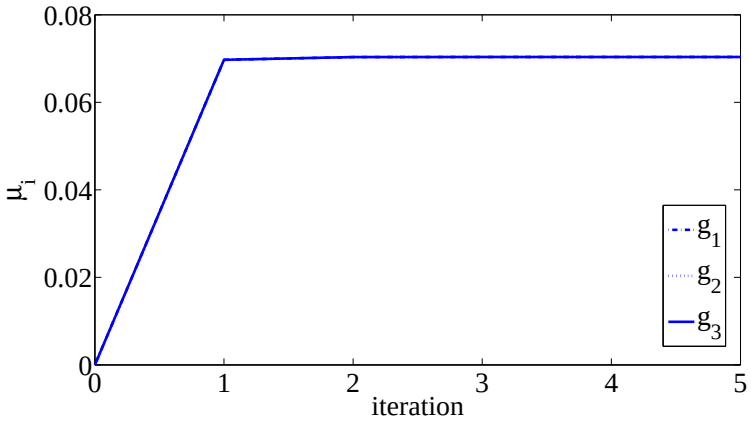

Figure 3.22: Convergence of g_i

numerical study, we assume each response follows uniform distribution between 20 and 40 minutes. All routing probabilities follow uniform distribution from 0 to 1. The time distribution of patient in normal status is exponential. Ten instances are generated. Since rapid response implies that most of the time the patients are in normal status, the ratio between declining and normal status for patients is always less than 20% in the setting. The accuracy results are shown in Table 3.5. Let $T_{final}^{sim,i}$ and $T_{final}^{sri,i}$ denote the mean decision time obtained by simulation and that obtained by two-level SRI in instance i. The average relative error, $\overline{\epsilon}$, shown below is defined as follows. The simulation time and number of trial run numbers are appropriately designed in simulation, so that the confidence limits are within 1% of the estimation of mean.

$$\overline{\epsilon} = \frac{\Sigma_{i=1}^{10} \frac{|T_{final}^{sim,i} - T_{final}^{sri,i}|}{T_{final}^{sim,i}}}{10} \cdot 100\%. \tag{3.122}$$

Table 3.5: Accuracy of two-level SRI, $\overline{\epsilon}$

Normal time(minutes)	300	350	400	450	500	550	600
2 patients	0.85%	0.90%	0.91%	0.90%	0.88%	0.86%	0.85%
3 patients	1.13%	1.06%	0.99%	0.96%	0.92%	0.88%	0.83%
4 patients	1.76%	1.65%	1.54%	1.44%	1.35%	1.29%	1.18%
5 patients	1.66%	1.55%	1.44%	1.30%	1.24%	1.19%	1.13%

From the above table, the two-level shared resource iteration performs well, which indicates its effectiveness.

3.6.4 Applicability

In the previous setting for accuracy test, exponential distribution is assumed for time duration of patient normal status. In most real cases, such an assumption may not hold. In this subsection, we investigate how the patient's normal time affects waiting due to resource sharing. To do this, gamma and lognormal distributions are used. Since in rapid response setting, the more time elapses, the more likely the patient will be at risk of deterioration. Therefore, we study the cases where CV is less than 1. Specifically, we take three data points, where CV = 0.25, 0.5, 0.75, as well as CV = 0 and CV = 1. We hypothesize that the variability has little effect on the new mean decision time . The simulation results match the hypothesis. In Table 3.6, ten instances are created and the largest possible relative error $\overline{\delta}$ is presented. Let $T_{i,j}$ denote within instance j, $j = 1, ..., 10$, the mean decision time obtained from simulation where $i = 1, 2, 3, 4$ corresponding to the cases where $CV = 0, 0.25, 0.5, 0.75, 1$, respectively.

$$\overline{\delta} = \frac{\sum_{j=1}^{10} \frac{|\max_i T_{i,j} - \min_i T_{i,j}|}{\min_i T_{i,j}}}{10} \cdot 100\%. \tag{3.123}$$

Table 3.6: Accuracy of two-level SRI in lognormal distribution case, $\overline{\delta}$

Normal time(minutes)	300	350	400	450	500	550	600
2 patients	0.07%	0.07%	0.07%	0.06%	0.08%	0.08%	0.07%
3 patients	0.24%	0.17%	0.14%	0.12%	0.14%	0.12%	0.13%
4 patients	0.37%	0.26%	0.25%	0.19%	0.17%	0.16%	0.20%
5 patients	0.46%	0.27%	0.21%	0.13%	0.12%	0.13%	0.82%

Similarly, as for gamma distribution, the results are shown in Table 3.7, which is also satisfactory to support the hypothesis. Therefore, we conclude that the proposed

iterative method can provide an accurate estimation of mean decision time in multiple patient scenario.

Table 3.7: Accuracy of two-level SRI in gamma distribution case, $\overline{\delta}$

Normal time(minutes)	300	350	400	450	500	550	600
2 patients	0.28%	0.22%	0.16%	0.14%	0.12%	0.11%	0.10%
3 patients	0.59%	0.42%	0.31%	0.21%	0.16%	0.14%	0.10%
4 patients	0.97%	0.70%	0.51%	0.41%	0.31%	0.27%	0.20%
5 patients	1.50%	1.03%	0.77%	0.64%	0.51%	0.38%	0.33%

3.7 Conclusions

In this chapter, a preliminary model to characterize the rapid response operations with dedicated RRT in acute care is developed. Closed formulas have been derived to evaluate the mean and variability of decision time. For response-time performance under exponential service time assumption, we derived a closed formula. In the study of extension to non-exponential case, we provide an approximation approach with satisfactory accuracy.

The monotonic properties of these performance measures have been investigated. Bottleneck identification methods are developed to improve the efficacy of rapid response operations in acute care. Specifically, the bottleneck response for decision time (BN-τ), decision time variability (BN-cv), and response-time performance (BN-rtp) have been introduced, and bottleneck indicators have been developed based on the collected data. Such methods provide a quantitative tool for analysis and improvement of rapid response

operations in acute care delivery to improve patient safety and care quality. Lastly, we address the resource sharing issue under the current framework. A two-level shared resource iteration approach is developed. Satisfactory accuracy is obtained and the applicability of the approach is discussed.

Chapter 4

Patient Rescue Process Framework and Modeling

4.1 Patient Rescue Process Framework

Improving patient safety is the top priority for hospital management. In recent years, reducing the rate of FTR has been a nationwide interest. One of the key issues is to identify hospital deteriorating patients and reduce preventable harm and mortality. For most FTR cases on hospital floors, patients may start with experiencing physiological deterioration, changing from non-risk condition to risk status. Such declining may trigger a signal to alarm or warn the nurse or may be detected by RNs frequent check. Possible intervention by the RN or physician may bring the patient back to normal condition. However, in some cases, RRT calls need to be initiated to request further treatment for rescue. Many RRTs are assembled by attending and fellow physicians from ICU. Such a process is typically referred to as patient rescue process, which is an integrated process with strong correlations and coordinations among multiple departments, and among different care providers. A systematical way to quantitatively model and improve this process is of importance.

In this chapter, a systems approach is presented to improve operations related to the

rescue process. An analytical framework with five modules is proposed to study the correlations and interactions among all the factors related to the process. More specifically, the five modules include: a triage module labels patients in different categories, such as moving to ICU or floor ward, risk, non-risk, etc.; a patient module describes the patient status on the floor, where possible deterioration may occur (i.e., status changing from non-risk to risk); a floor module represents the RN check and possible intervention by providers (RN and MD), and initiation of RRT calls; a RRT module describes the effort and treatment by RRT; since many RRTs are assembled by physicians from ICU, thus an ICU module outlines the staffing demand in ICU practices. Such a framework provides an integrated model to address the whole process systematically. An illustration of such a framework is shown in Figure 4.1.

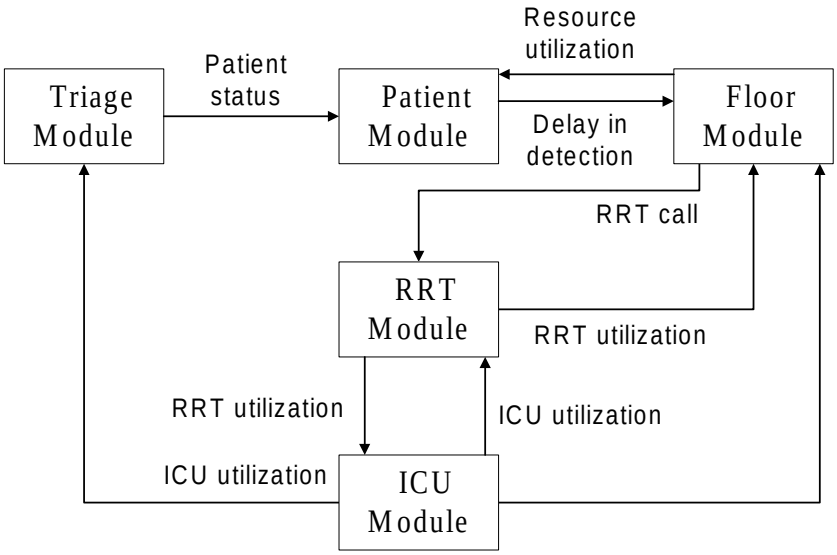

Figure 4.1: Analytical framework of rescue process

As one can see, the triage module identifies admitted patient status, which feeds into the patient module to generate declining signals. Such declining may trigger a signal to alarm the RN or may be detected by the frequent check of RN in the floor module.

RRT calls may be initiated if severe and continuous deterioration is identified. Clearly, the RRT utilization will affect ICU operations. However, the ICU utilization may also impact RRT deployment, and may change triage decision if limited capacity is observed. To understand the complex interactions among these modules and their implications, quantitative models need to be developed to study the impacts of different factors in these modules. Stochastic models or discrete event simulation models can be developed to study the patient rescue process from different perspectives.

4.2 Patient Rescue Process Modeling

Using the framework described above, we study the patient rescue process to patient deterioration. This includes two most critical phases: the first phase is the detection and recognition of patient deterioration, which is indicated by abnormal physiological vital signs, such as heart rate, respiratory rate, and blood pressure. The second phase is the intervention after deterioration happens. Multiple care providers, such as the nurses (RN), physicians (MD), rapid response team (RRT), and/or staffs from intensive care unit (ICU) are involved in the process Berwick et al. [13], Priestley et al. [88]. Time delay in either of the two phases will highly likely put patient safety at risk McGloin et al. [80] and impose extra burden on ICU McArthur-Rouse [78].

Since the patient rescue process is a complex process involving multiple providers in different disciplinary or divisions and various patient syndromes or complications, there is a need to investigate the problem from a system vista. A system model can provide a fresh look at the problem by developing a novel investigation approach to study the dynamics and interactions through quantitative modeling of the rescue process. We introduce an analytical model to analyze such process in the hospital, which can address

the interactions between patient condition change and staff response, such as RN or MD intervention, RRT call initiation, etc. It can also provide insights to aid decision making on staffing allocation, team composition, detection and intervention protocols, etc.

In this study, a continuous time Markov chain model is developed,which characterizes the patient status as normal (non-risk), deteriorating (risk), being intervened by RN, MD and RRT, and being elevated (to ICU). Closed formulas to calculate the probability of the patient staying in different states are developed for single patient case. An approximation method, referred to as the shared resource iteration (SRI) approach, is proposed to study the multiple patients scenario. Using numerical study, such an iteration is found to be convergent and results in a high accuracy estimation of patient state probability. This method provides a quantitative tool to analyze the patient rescue process and investigate strategies to improve patient safety. The successful development of such a model can provide hospital management a quantitative tool to analyze the efficacy of patient rescue process and seek recommendations for improvement.The detailed model and analysis method are introduced below.

4.2.1 System Description and Assumptions

System Description

A typical patient rescue process is shown in Figure 4.2 and described as follows:

- The inpatient is continuously monitored by an alarm system and through regular check by the RN, who works as the first line provider in most hospitals. If the patient is in normal status, no intervention is needed and the patient will be checked back in next round.

- The condition of the patient on the floor can be classified into two categories:

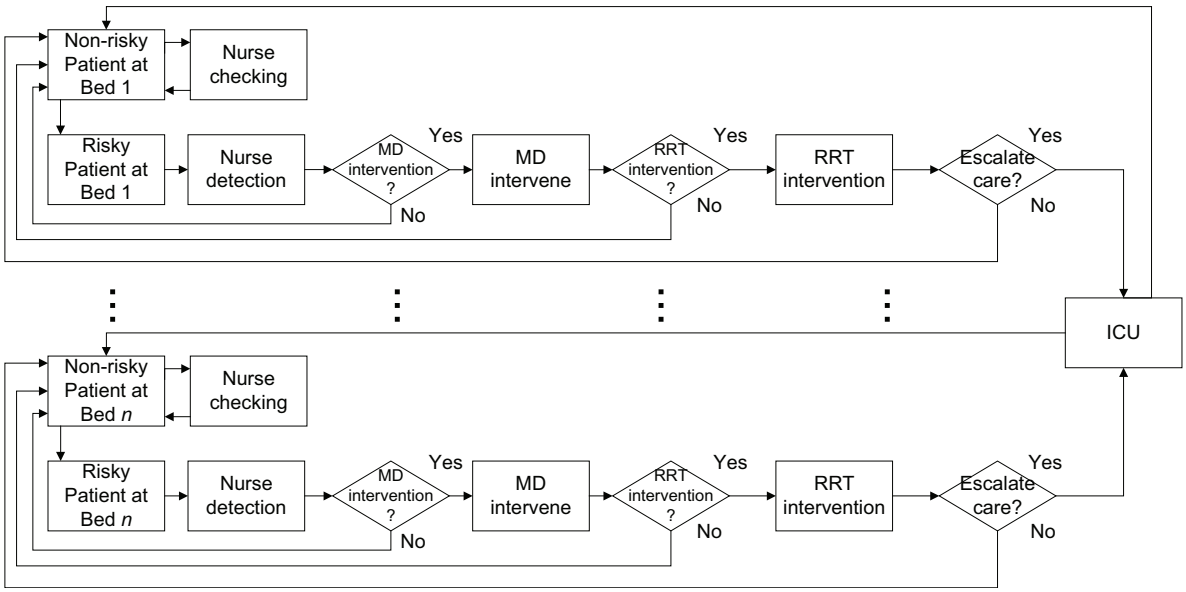

Figure 4.2: Flow diagram of patient rescue process

"non-risk" implies that the patient is free of clinical deterioration, while "risk" indicates that the patient is susceptible to acute physiological deterioration, which is a precursor to serious and often fatal outcomes, such as cardiac arrest.

- The patient may experience physiological deterioration, changing from non-risk condition to risk status.

- The patient declining may trigger a signal to alarm or warn the nurse (RN) or may be detected by RN's frequent check.

- If the patient's deterioration is not severe, the RN's intervention may bring the patient back to non-risk condition. However, if the declining is more critical, the MD should be notified.

- In some cases, the MD's intervention can bring the patient back to normal condition. However, in some other cases, a RRT call need to be initiated to request

further treatment for rescue.

- Again, the RRT treatment can either bring the patient back to normal, or suggest further rescue, i.e., elevating the patient for ICU admission.

- The patient elevated to ICU may become stable after treatment, and is de-elevated back to the floor to normal condition. Hospital mortality is not considered in this model.

As one can see, patient rescue is an integrated process with strong correlations and coordinations among different departments, and among many care providers. A systematical way to quantitatively model and improve this process is of importance.

Assumptions

To study such a process, the following assumptions addressing the patient's status, provider's intervention, and their correlations, are introduced:

1) There are n inpatients on the hospital floor.

2) The care provider team is composed of one RN, one MD, and one RRT.

3) For a non-risky patient i, $i = 1, \ldots, n$, the time between RN's regular checked is exponentially distributed with parameter $\lambda_{i,1}$, which is referred to as the nurse check rate.

4) The time a normal patient i, $i = 1, \ldots, n$, stays at non-risk condition is exponentially distributed with parameter $\lambda_{i,2}$. Such a parameter is also referred to as a patient's declining rate.

5) For a risky patient i, $i = 1, \ldots, n$, the time until a RN detects the patient's declining is described by an exponential distribution with parameter $\lambda_{i,3}$. This parameter is defined as the detection rate.

6) For a non-risky patient i, $i = 1, \ldots, n$, the time the RN needs to spend for a regular check follows exponential distribution with parameter $\mu_{i,1}$.

7) For a risky patient i, $i = 1, \ldots, n$, the time from intervention by RN, MD, and RRT to back to normal are exponentially distributed with parameters $\mu_{i,2}$, $\mu_{i,3}$, and $\mu_{i,4}$, respectively. These parameters are also referred to as the RN, MD, and RRT's rescue rates, respectively.

8) The ICU treatment time for an elevated patient i, $i = 1, \ldots, n$, is also exponentially distributed with parameter $\mu_{i,5}$, which is referred to as the recover rate. It is assumed that the ICU has enough capacity to admit patients if needed.

9) For a risky patient i, $i = 1, \ldots, n$, the time from intervention by RN, MD, and RRT to the arrival of upper level providers (MD, RRT, and ICU, respectively) are exponentially distributed with $\lambda_{i,4}$, $\lambda_{i,5}$ and $\lambda_{i,6}$, respectively.

The justification of the above assumptions are discussed in the following remarks.

Remark 4.1 The exponential assumptions are introduced to make the analysis tractable. Such assumptions can be relaxed in future work.

Remark 4.2 Assumption 9) implies that, to intervene a risky patient, the responding provider will wait until the higher level one arrives. In other words, the RN will stay with the patient until the MD is coming, or the MD will stay until the RRT arrives.

Remark 4.3 According to the above assumptions, self-healing is not considered for deteriorating patients in acute care.

4.2.2 Single Patient Case

Problem Formulation

We first consider single patient scenario. In an appropriately defined state space, the patient rescue process described above is a stationary random process. In the framework of 1)-9), a continuous-time Markov chain (CTMC) model can be developed. Introduce Markov states S_1 to S_7 to denote the status of the patient, representing risk, non-risk, provider intervention, etc. These states are defined as follows:

- State S_1: "Floor NR" - the patient is in non-risk condition, waiting to be checked by the RN.
- State S_2: "Floor R" - the patient is in risk condition, waiting to be checked by the RN.
- State S_3: "Nurse NR" - the patient is in non-risk condition and the RN is checking him/her.
- State S_4: "Nurse R" - the patient is in risk condition and under RN intervention.
- State S_5: "MD Int" - the patient is in deteriorating condition and under MD intervention.
- State S_6: "RRT Int" - the patient is in deteriorating condition and under RRT intervention.
- State S_7: "ICU" - the patient is in deteriorating condition and is elevated and treated in ICU.

Such states and transitions for the patient are illustrated in Figure 4.3.

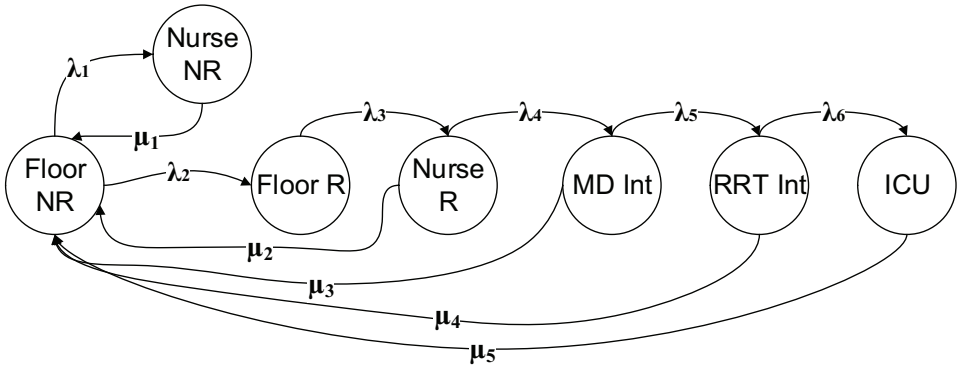

Figure 4.3: CTMC transition diagram for one patient

Let P_i, $i = 1, \ldots, 7$, denote the steady state probability that the patient is in state i. In the framework of 1)-9), P_i is a function of all system parameters, i.e.,

$$P_i = f_{P,i}(\Lambda, \Gamma), \quad i = 1, \ldots, 7, \tag{4.1}$$

where

$$\begin{aligned} \Lambda &= [\lambda_1, \ldots, \lambda_6], \\ \Gamma &= [\mu_1, \ldots, \mu_5]. \end{aligned}$$

Solutions to the problem are presented below.

Performance Evaluation

Using the CTMC model under single-patient case, the balance equations can be written as follows:

$$(\lambda_1 + \lambda_2)P_1 = \mu_1 P_3 + \mu_2 P_4 + \mu_3 P_5 + \mu_4 P_6 + \mu_5 P_7, \tag{4.2}$$

$$\lambda_3 P_2 = \lambda_2 P_1, \tag{4.3}$$

$$\mu_1 P_3 = \lambda_1 P_1, \tag{4.4}$$

$$(\lambda_4 + \mu_2)P_4 = \lambda_3 P_2, \tag{4.5}$$

$$(\lambda_5 + \mu_3)P_5 = \lambda_4 P_4, \tag{4.6}$$

$$(\lambda_6 + \mu_4)P_6 = \lambda_5 P_5, \tag{4.7}$$

$$\mu_5 P_7 = \lambda_6 P_6. \tag{4.8}$$

The steady state probabilities can be solved. Since the Markov chain under study is finite, irreducible and positive recurrent, the limiting probabilities will exist. The steady-state probabilities can be calculated as follows:

$$P_1 = \frac{1}{\Psi}, \tag{4.9}$$

$$P_2 = C_2 P_1, \tag{4.10}$$

$$P_3 = C_3 P_1, \tag{4.11}$$

$$P_4 = C_2 C_4 P_1, \tag{4.12}$$

$$P_5 = C_2 C_4 C_5 P_1, \tag{4.13}$$

$$P_6 = C_2 C_4 C_5 C_6 P_1, \tag{4.14}$$

$$P_7 = C_2 C_4 C_5 C_6 C_7 P_1, \tag{4.15}$$

where,

$$\Psi = C_1 + C_2 + C_3 + C_2 C_4 + C_2 C_4 C_5 + C_2 C_4 C_5 C_6 + C_2 C_4 C_5 C_6 C_7,$$

$$C_1 = 1,$$

$$C_2 = \lambda_2/\lambda_3,$$

$$C_3 = \lambda_1/\mu_1,$$

$$C_4 = \lambda_3/(\lambda_4 + \mu_2),$$

$$C_5 = \lambda_4/(\lambda_5 + \mu_3),$$

$$C_6 = \lambda_5/(\lambda_6 + \mu_4),$$

$$C_7 = \lambda_6/\mu_5.$$

The above results are obtained from balance equations (4.2) to (4.8), we can rewrite P_i as:

$$\begin{aligned}
P_2 &= \frac{\lambda_2}{\lambda_3} P_1 = C_2 P_1, \\
P_3 &= \frac{\lambda_1}{\mu_1} P_1 = C_3 P_1, \\
P_4 &= \frac{\lambda_3}{\lambda_4 + \mu_2} P_2 = C_4 P_2 = C_4 C_2 P_1, \\
P_5 &= \frac{\lambda_4}{\lambda_5 + \mu_3} P_4 = C_5 P_4 = C_5 C_4 C_2 P_1, \\
P_6 &= \frac{\lambda_5}{\lambda_6 + \mu_4} P_5 = C_6 P_5 = C_6 C_5 C_4 C_2 P_1, \\
P_7 &= \frac{\lambda_6}{\mu_5} P_6 = C_7 P_6 = C_7 C_6 C_5 C_4 C_2 P_1.
\end{aligned}$$

From the normalization condition that the summation of all P_is equal to 1, we have

$$\begin{aligned}
&P_1(1 + C_2 + C_3 + C_4 C_2 + C_5 C_4 C_2 + C_6 C_5 C_4 C_2 \\
&\quad + C_7 C_6 C_5 C_4 C_2) = 1.
\end{aligned}$$

Solve for P_1, then all other P_is can be obtained.

Furthermore, the utilizations of the RN, MD and RRT can be obtained as well by summing up all the corresponding probabilities. Specifically, under assumptions 1)-9), for single patient case, the utilizations of RN, MD, and RRT are calculated as follows:

$$\rho_{RN} = P_2 + P_4, \tag{4.16}$$

$$\rho_{MD} = P_5, \tag{4.17}$$

$$\rho_{RRT} = P_6. \tag{4.18}$$

4.2.3 Multiple Patients Case

The above results are only applicable when only one patient is considered, in which the providers are always available to carry out the service. When multiple patients are involved, due to limited resource, such services may not be available. In other words, a patient may need to wait if the requested provider is serving another patient. For example, as shown in Figure 4.4, Nurse can be shared by multiple patients, represented by the large rectangles. Similar scenarios can be observed for MD and RRT as well.

Intuitively, one may define and add new states to the existing CTMC model to address this issue. For example, one can explicitly list different combinations of providers' services on different patients as states. However, the relationships between the states will be extremely complex for analysis because we have to consider differentiated patients, a variety of resources, each with a different number, as well as numerous waiting scenarios. Indeed, under the two-identical-patient case, a CTMC model with 29 states can be developed. Such analytical model is discussed below.

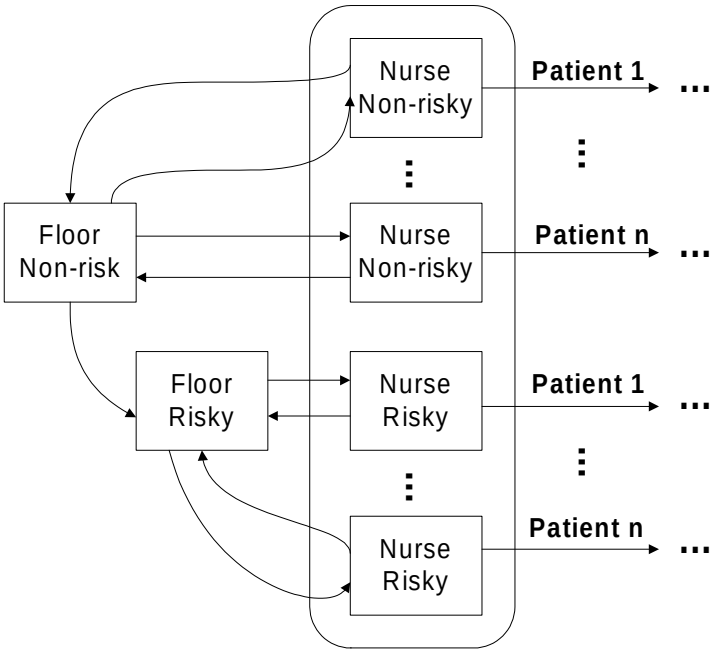

Figure 4.4: Illustration of provider resource sharing

A two-identical-patient analytical model

Consider a two-identical-patient case, where transition rates are identical (i.e., $\lambda_{k,i} = \lambda_i, \mu_{k,i} = \mu_i, k = 1, 2.$), with one set of providers (1 RN, 1 MD, 1 RRT). A CTMC model can be developed. Each state in the two-identical-patient CTMC model is in the form of a vector:

$[A_{FloorNR}, A_{FloorR}, A_{NurseNR}, A_{NurseR}, A_{MDInt}, A_{RRTInt}, A_{ICU}]$, defined as follows:

- $A_{FloorNR}$: number of patients in non-risk condition.
- A_{FloorR}: number of patients in risk condition.
- $A_{NurseNR}$: number of patients in non-risk condition and the RN is checking him/her.
- A_{NurseR}: number of patients in risk condition and under RN intervention.

- A_{MDInt}: number of patients in deteriorating condition and under MD intervention.
- A_{RRTInt}: the number of patients in deteriorating condition and under RRT intervention.
- A_{ICU}: number of patients in deteriorating condition and is elevated and treated in ICU.

Moreover, for elements $A_{FloorNR}, A_{FloorR}, A_{NurseR}, A_{MDInt}$, let 1b denote one patient is in the respective status but being blocked, i.e., waiting. We list all the Markov states of the two-identical-patient CTMC model below with the numbering:

1: (0,0,0,0,0,0,2)	2: (0,0,0,0,0,1,1)	3: (0,0,0,0,1,0,1)	4: (0,0,0,1,0,0,1)
5: (0,0,1,0,0,0,1)	6: (0,1,0,0,0,0,1)	7: (1,0,0,0,0,0,1)	8: (0,0,0,0,1,1,0)
9: (0,0,0,0,1b,1,0)	10: (0,0,0,1,0,1,0)	11: (0,0,1,0,0,1,0)	12: (0,1,0,0,0,1,0)
13: (1,0,0,0,0,1,0)	14: (0,0,0,1,1,0,0)	15: (0,0,0,1b,1,0,0)	16: (0,0,1,0,1,0,0)
17: (0,1,0,0,1,0,0)	18: (1,0,0,0,1,0,0)	19: (0,1,0,1,0,0,0)	20: (0,1b,0,1,0,0,0)
21: (1,0,0,1,0,0,0)	22: (1b,0,0,1,0,0,0)	23: (0,1,1,0,0,0,0)	24: (0,1b,1,0,0,0,0)
25: (1,0,1,0,0,0,0)	26: (1b,0,1,0,0,0,0)	27: (0,2,0,0,0,0,0)	28: (1,1,0,0,0,0,0)
29: (2,0,0,0,0,0,0)			

For example, state 1 means two patients are in ICU. State 2 implies that 1 patient is in ICU, while the other one is under RRT intervention. State 9 is the state with blockage: The MD has delivered care to one patient and ask for RRT intervention, but that patient has to wait for RRT, which is with the other patient. The transition rate matrix Q is presented in Appendix B.

Since the two-identical-patient Markov chain is finite, irreducible and positive recurrent, the limiting probabilities exist. Let π_i denote the steady state probability of state i. From balance equations and normalizing condition, we can obtain π_i, $i = 1, 2, \ldots, 29$.

Using such π_is, we can obtain the steady state probability of the two-identical-patient CTMC model.

In case of more than two patients being considered, there can be an exponential increase in number of states when the number of patients increases, which makes directly list all the states impossible. Therefore, we introduce the idea of the iterative method, we investigate the performance measures of this two-identical-patient system by approximation using the single-patient Markov state introduced earlier. Specifically, we construct the mapping of the approximation as follows.

$$
\begin{aligned}
2P_1 &\approx \pi_7 + \pi_{13} + \pi_{18} + \pi_{21} + \pi_{22} + \pi_{25} + \pi_{26} + \pi_{28} + \pi_{29}, \\
2P_2 &\approx \pi_6 + \pi_{12} + \pi_{17} + \pi_{19} + \pi_{20} + \pi_{23} + \pi_{24} + \pi_{27} + \pi_{28}, \\
2P_3 &\approx \pi_5 + \pi_{11} + \pi_{16} + \sum_{k=23}^{26} \pi_k, \\
2P_4 &\approx \pi_4 + \pi_{10} + \pi_{14} + \pi_{15} + \sum_{k=19}^{22} \pi_k, \\
2P_5 &\approx \pi_3 + \pi_8 + \pi_9 + \sum_{k=14}^{18} \pi_k, \\
2P_6 &\approx \pi_2 + \sum_{k=8}^{13} \pi_k, \\
2P_7 &\approx \sum_{k=1}^{7} \pi_k.
\end{aligned}
$$

In the above approximation, the left hand side equals to the room number multiply by steady state probability of single-patient Markov model, whose closed formula can be obtained. The right hand side equals the corresponding probability that one patient is in state i in the two-identical-patient Markov model. The error in approximation is that first, the Markovian assumption is used for both service time and waiting time. Second, the combination of π_is does not always reflect the true respective performance measure.

For example, in the first approximation, which uses $2P_1$, the true probability that one patient is in non-risk condition, should equal to $\pi_7 + \pi_{13} + \pi_{18} + \pi_{21} + \pi_{22} + \pi_{25} + \pi_{26} + \pi_{28} + 2\pi_{29}$.

For identical patient case where number of patient is greater than 2, similar idea can be developed to use Markov states in a single-patient model to approximate the respective performance measures of system with n patients. The illustration of the method is presented below.

An Illustrative Example

setting with one set of providers (1 RN, 1 MD, 1 RRT) focusing on RN occupancy is introduced. Let $\lambda_{k,1}^{(j)}$ and $\lambda_{k,3}^{(j)}$ denote the RN check rate and detection rate for patient k, $k = 1, 2$, during the j-th iteration, $j = 1, 2, \ldots$, respectively. Introduce $P_{k,3}^{(j)}$ and $P_{k,4}^{(j)}$ to represent the steady state probabilities of state *Nurse NR* and state *Nurse R* for patient k at iteration j, respectively.

To start the iteration, first we assume the RN's occupancy in patient ward k when a patient is in risky or non-risky condition is known. In other words, $P_{k,3}^{(0)}$ and $P_{k,4}^{(0)}$, $k = 1, 2$, are known (for instance, they can be calculated using single patient assumption).

Now we consider patient 1. If the RN is currently serving patient 2, then he/she has to wait until the RN is available. Thus, patient 1's status can only change when the RN service to the other patient is finished. To represent such a scenario, we update the RN check and detection rates as follows:

$$\lambda_{1,1}^{(1)} = \lambda_{1,1}^{(0)}\Big(1 - P_{2,3}^{(0)} - P_{2,4}^{(0)}\Big),$$
$$\lambda_{1,3}^{(1)} = \lambda_{1,3}^{(0)}\Big(1 - P_{2,3}^{(0)} - P_{2,4}^{(0)}\Big),$$

where $\lambda_{1,1}^{(0)} = \lambda_{1,1}$ and $\lambda_{1,3}^{(0)} = \lambda_{1,3}$. In other words, we assume the transition rates are

reduced due to resource sharing from other patients.

Using the new RN check and detection rates, we can re-evaluate the RN service for patient 1, i.e., an update of $P_{1,2}$ and $P_{1,4}$ can be obtained using the closed formulas in single patient case.

$$\begin{aligned} P_{1,2}^{(1)} &= f_{P,2}\Big(\Lambda_1^{(1)}, \Gamma_1\Big), \\ P_{1,4}^{(1)} &= f_{P,4}\Big(\Lambda_1^{(1)}, \Gamma_1\Big), \end{aligned}$$

where $\lambda_{1,1}^{(1)}$ and $\lambda_{1,3}^{(1)}$ are used to replace $\lambda_{1,1}^{(0)}$ and $\lambda_{1,3}^{(0)}$ in $\Lambda_1^{(1)}$, respectively.

Now consider patient 2. Similar to the analysis of patient 1, the nurse check and detection rates are updated.

$$\begin{aligned} \lambda_{2,1}^{(1)} &= \lambda_{2,1}^{(0)}\Big(1 - P_{1,3}^{(1)} - P_{1,4}^{(1)}\Big), \\ \lambda_{2,3}^{(1)} &= \lambda_{2,3}^{(0)}\Big(1 - P_{1,3}^{(1)} - P_{1,4}^{(1)}\Big), \end{aligned}$$

and $P_{2,2}$ and $P_{2,4}$ will be updated, i.e.,

$$\begin{aligned} P_{2,2}^{(1)} &= f_{P,2}\Big(\Lambda_2^{(1)}, \Gamma_2\Big), \\ P_{2,4}^{(1)} &= f_{P,4}\Big(\Lambda_2^{(1)}, \Gamma_2\Big), \end{aligned}$$

where $\lambda_{2,1}^{(1)}$ and $\lambda_{2,3}^{(1)}$ are used in $\Lambda_2^{(1)}$.

After all the steady state probabilities are updated, we finish the first iteration. Using these results, we start the second iteration, and repeat such a process until the differences of steady-state probabilities between two successive iterations are small enough, i.e., the process converges. Then the steady-state probabilities of patient states are obtained.

If the MD and RRT interventions are also shared by multiple patients, similar treatment can be applied to calculate their steady state probabilities iteratively. Notice that all the steady-state probabilities, $P_{k,i}$s, are approximations, which is the combination of

steady-state probabilities of the n-patient CTMC model. These $P_{k,i}$s can provide the utilization of resources, which does not necessitate knowing each specific individual state probabilities of n-patient CTMC model, which is extremely hard to model.

General Procedure

Using the idea introduced in the illustrative example, a general iterative procedure has been developed to approximate aggregated steady state probabilities of patient states. Consider a general patient rescue process with n patients, m_n RNs, m_d MDs, and m_t RRTs, where all m_n, m_d, $m_r < n$. Introduce operator $\Phi_i(n, m, P_{1,i}, \ldots, P_{n,i})$ to denote the probability that m providers are taking care of the patients in process i among n wards, and $P_{k,i}$, $k = 1, \ldots, n$, represents the probability that patient k is in process i, being served by the provider. Thus, operator $\Phi_i(\cdot)$ can be obtained as follows:

$$\begin{aligned}
&\Phi_i(n, m, P_{1,i}, \ldots, P_{n,i}) \\
&\quad = \sum_{k=1}^{n-m} P_{k,i} \Phi_i(n-k, m-1, P_{k+1,i}, \ldots, P_{n,i}) \\
&\qquad + \prod_{k=n-m+1}^{n} P_{k,i}, \qquad (4.19)
\end{aligned}$$

with initial condition

$$\Phi_i(l, 1, P_{1,i}, \ldots, P_{l,i}) \;=\; \sum_{k=1}^{l} P_{k,i}, \qquad (4.20)$$

and boundary condition

$$\Phi_i(m, m, P_{1,i}, \ldots, P_{m,i}) \;=\; \prod_{k=1}^{m} P_{k,i}. \qquad (4.21)$$

When patient k is in process i waiting for the providers, who are serving patients in other wards in this process, the probability of such an event is described by $\Phi_i(n -$

$1, m, P_{1,i}, \ldots, P_{k-1,i}, P_{k+1,i}, \ldots, P_{n,i})$, where $m = m_n$ if it is in RN service (respectively, $m = m_d$ or m_r for MD and RRT services).

Therefore, the general iterative procedure for a patient rescue process with multiple patients and providers can be described as follows:

Procedure 4.1 Step 1: Initialization: *Calculate steady-state probabilities under single patient setting and set those as initial values.*

$$P_{k,i}^{(0)} = f_{P,i}(\Lambda_k, \Gamma_k), \quad i = 1, \ldots, 7, k = 1, \ldots, n.$$

Step 2: Patient update: *Update transition rates related to the occupancy of RN, MD and RRT for patient k, $k = 1, \ldots, n$, and obtain updated steady-state probabilities.*

During iteration j, $j = 0, 1, 2, \ldots$,

$$\begin{aligned}
\lambda_{k,1}^{(j+1)} &= \lambda_{k,1}^{(0)}\Big[1 - \Phi_3\Big(n-1, m_n, P_{1,3}^{(j)}, \ldots, P_{k-1,3}^{(j)}, \\
&\qquad P_{k+1,3}^{(j)}, \ldots, P_{n,3}^{(j)}\Big) - \Phi_4\Big(n-1, m_n, P_{1,4}^{(j)}, \\
&\qquad \ldots, P_{k-1,4}^{(j)}, P_{k+1,4}^{(j)}, \ldots, P_{n,4}^{(j)}\Big)\Big], \\
\lambda_{k,3}^{(j+1)} &= \lambda_{k,3}^{(0)}\Big[1 - \Phi_3\Big(n-1, m_n, P_{1,3}^{(j)}, \ldots, P_{k-1,3}^{(j)}, \\
&\qquad P_{k+1,3}^{(j)}, \ldots, P_{n,3}^{(j)}\Big) - \Phi_4\Big(n-1, m_n, P_{1,4}^{(j)}, \\
&\qquad \ldots, P_{k-1,4}^{(j)}, P_{k+1,4}^{(j)}, \ldots, P_{n,4}^{(j)}\Big)\Big], \\
\lambda_{k,4}^{(j+1)} &= \lambda_{k,4}^{(0)}\Big[1 - \Phi_5\Big(n-1, m_d, P_{1,5}^{(j)}, \ldots, P_{k-1,5}^{(j)}, \\
&\qquad P_{k+1,5}^{(j)}, \ldots, P_{n,5}^{(j)}\Big)\Big], \\
\lambda_{k,5}^{(j+1)} &= \lambda_{k,5}^{(0)}\Big[1 - \Phi_6\Big(n-1, m_r, P_{1,6}^{(j)}, \ldots, P_{k-1,6}^{(j)}, \\
&\qquad P_{k+1,6}^{(j)}, \ldots, P_{n,6}^{(j)}\Big)\Big], \\
P_{k,i}^{(j+1)} &= f_{P,i}\Big(\Lambda_k^{(j+1)}, \Gamma_k\Big), \quad i = 2, 4, 5, 6.
\end{aligned}$$

Step 3: Iteration: *Let $j = j + 1$, go back to Step 3 until the stopping criteria is met. For a given $\delta \ll 1$, the iteration can stop when all the differences in $P_{k,i}$ between two consecutive iterations are smaller than δ, i.e.,*

$$\max_{k,i} \left| P_{k,i}^{(j)} - P_{k,i}^{(j-1)} \right| < \delta, \forall k = 1, \ldots, n, i = 3, \ldots, 6.$$

Next, the convergence of such an iterative method is studied. Proposition 4.1 and Proposition 4.2 are presented below.

Proposition 4.1 *For Procedure 4.1, if $n = 2$, $m_d = 1$, $m_n = m_r = 2$, the following limits exist:*

$$\lim_{j \to \infty} P_{k,5}^{(j)} = P_{k,5}, k = 1, 2, \tag{4.22}$$

$$\lim_{j\to\infty} P_{k,6}^{(j)} = P_{k,6}, k = 1, 2. \tag{4.23}$$

Proof: See Appendix A. ■

Proposition 4.2 *For Procedure 4.1, if* $n = 2$, $m_r = 1$, $m_n = m_d = 2$, *the following limits exist:*

$$\lim_{j\to\infty} P_{k,5}^{(j)} = P_{k,5}, k = 1, 2, \tag{4.24}$$

$$\lim_{j\to\infty} P_{k,6}^{(j)} = P_{k,6}, k = 1, 2. \tag{4.25}$$

Proof: Analogously to the proof of Proposition 4.1. ■

The above propositions state that, for a 2-patient system, both $P_{k,5}$ and $P_{k,6}$, $k = 1, 2$ are convergent if number of MD or RRT equals to 1. Therefore the utilization of MD and RRT can be estimated if only one of those two resource, can be shared.

For system with patient number greater than 2, and resource sharing is possible for RN, MD and RRT. The convergence property is investigated numerically. Based on extensive numerical experiments by randomly selecting number of patients, numbers of providers, and service time parameters, we justify that this iterative approach is convergent and results in a unique solution (i.e., independent of initial values). Therefore, we formulate this as a numerical fact:

Numerical Fact 4.1 *: Under assumptions 1)-9), Procedure 4.1 is convergent, i.e.,*

$$\lim_{j\to\infty} P_{k,i}^{(j)} = P_{k,i}, \forall k = 1, \ldots, n, i = 1, \ldots, 7. \tag{4.26}$$

In addition, the steady state probability solution, $P_{k,i}$, $k = 1, \ldots, n$, $i = 1, \ldots, 7$, *is unique.*

Using the convergent probabilities, the providers' utilizations are also obtained.

Corollary 4.1 *Under assumptions 1)-9), the utilizations of RN, MD, and RRT are calculated as follows:*

$$\rho_{RN} = \sum_{k=1}^{n}(P_{k,2} + P_{k,4}), \tag{4.27}$$

$$\rho_{MD} = \sum_{k=1}^{n} P_{k,5}, \tag{4.28}$$

$$\rho_{RRT} = \sum_{k=1}^{n} P_{k,6}. \tag{4.29}$$

The figures 4.5, 4.6, 4.7 and 4.8 illustrate the aforementioned numerical fact.

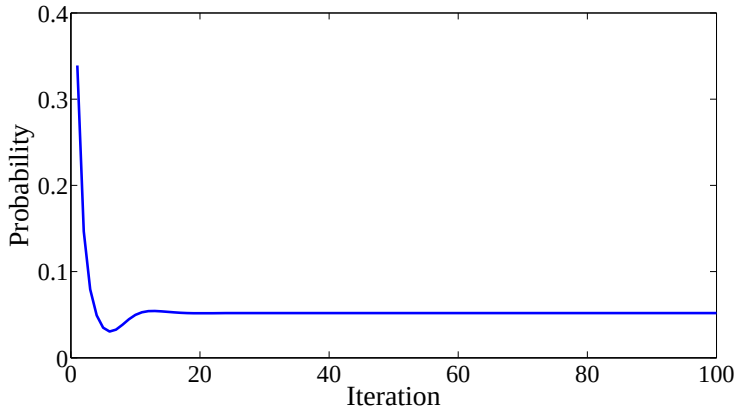

Figure 4.5: Probability of "Nurse NR" state for patient 1 ($P_{1,3}$)

4.2.4 Accuracy Investigation

In numerous settings, the accuracy of the method is investigated numerically by comparing with simulations. Dozens of examples have been generated with parameters selected randomly for multiple patients scenarios. In all simulations, we set the simulation time

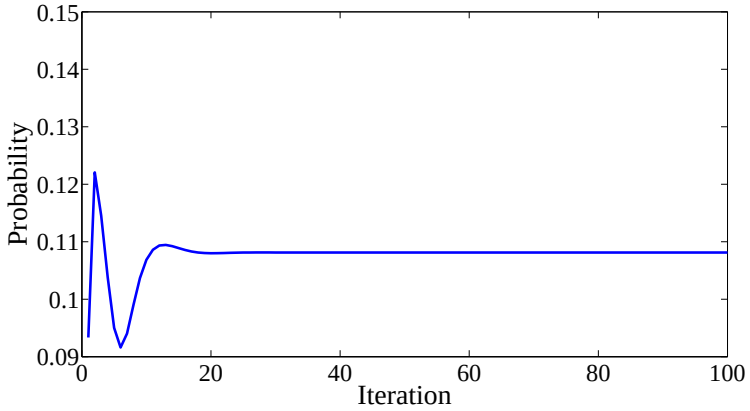

Figure 4.6: Probability of "Nurse R" state for patient 1 ($P_{1,4}$)

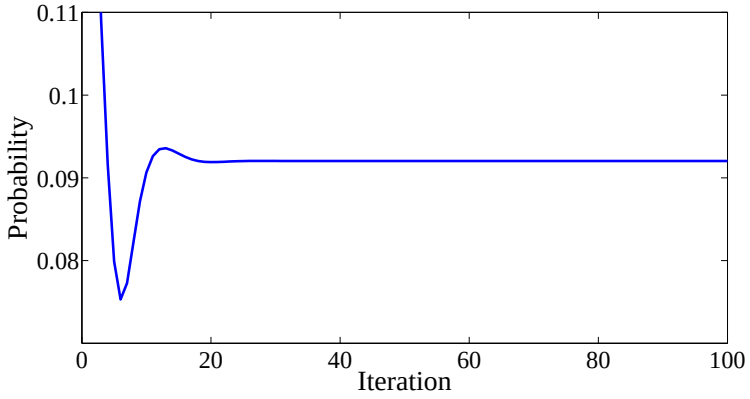

Figure 4.7: Probability of "MD Int" state for patient 1 ($P_{1,5}$)

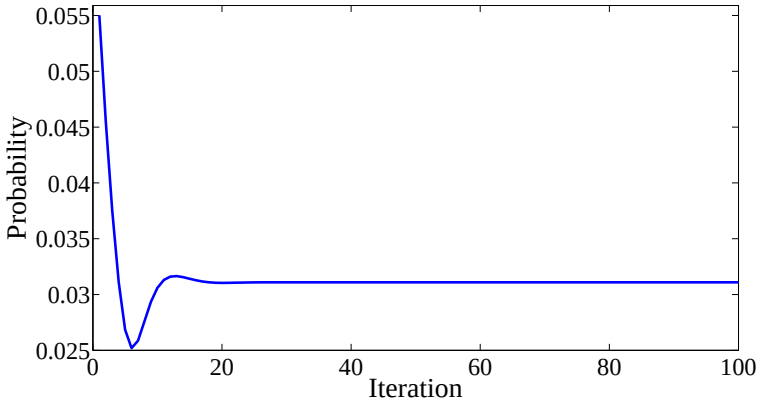

Figure 4.8: Probability of of "RRT Int" state for patient 1 ($P_{1,6}$)

and number of trial run numbers to make sure that the confidence limits are within 1% of the estimation of mean.

The average values of λ_i and μ_i are estimated by medical professionals working on hospital floor. Such estimates are shown in Tables 4.1 and 4.2 for λ_i and μ_i, respectively.

Table 4.1: Estimates of average λ_i (1/hour)

$\lambda_{est,1}$	$\lambda_{est,2}$	$\lambda_{est,3}$	$\lambda_{est,4}$	$\lambda_{est,5}$	$\lambda_{est,6}$
0.33	0.2	1	2.5	2	3.33

Table 4.2: Estimates of average μ (1/hour)

$\mu_{est,1}$	$\mu_{est,2}$	$\mu_{est,3}$	$\mu_{est,4}$	$\mu_{est,5}$
5	3.33	2.5	2	0.2

To accommodate the randomness of providers' services, parameters $\lambda_{k,i}$, $i = 1, \ldots, 6$,

$k = 1, \ldots, n$, and $\mu_{k,i}$, $i = 1, \ldots, 5$, $k = 1, \ldots, n$, are generated from the intervals

$$[\lambda_{est,i}/1.5, \lambda_{est,i} \cdot 1.5], \quad i = 1, \ldots, 6,$$
$$[\mu_{est,i}/1.5, \mu_{est,i} \cdot 1.5], \quad i = 1, \ldots, 5, \tag{4.30}$$

randomly with equal probability, respectively.

First, we consider the cases of identical patients (i.e., all wards have equal probability to accept any patients so that identical parameters are obtained for each ward, $\lambda_{k,i} = \lambda_i, \mu_{k,i} = \mu_i, \forall k$) are studied. Let ρ_{RN}^{sim}, ρ_{MD}^{sim}, and ρ_{RRT}^{sim} denote the utilizations of RN, MD, and RRT, obtained by simulations, respectively. Similarly, ρ_{RN}^{sri}, ρ_{MD}^{sri}, and ρ_{RRT}^{sri} represents the utilizations obtained using the SRI method. The relative error in utilization evaluation between simulation and iterative approach is defined in Equation (4.31):

$$\delta_i = \frac{\left|\rho_i^{sim} - \rho_i^{sri}\right| \cdot 100\%}{\rho_i^{sim}}, \quad i = RN, MD, RRT, \tag{4.31}$$

Because the utilizations of resources are very small numbers, to avoid small denominators when they are very small, the absolute error in utilization evaluation between simulation and iterative approach is defined in Equation (4.32):

$$\epsilon_i = \left|\rho_i^{sim} - \rho_i^{sri}\right|\cdot, \quad i = RN, MD, RRT. \tag{4.32}$$

where ρ_{RN}, ρ_{MD}, and ρ_{RRT} are defined in Corollary 4.1.

In each simulation experiment, 1000 hours of simulation time are assumed and 20 replications are carried out, to ensure steady state and sufficiently small confidence intervals. The number of patients is selected as 2, 4, 6, or 8. For each setting of providers, 10 simulation experiments by randomly selecting parameters from sets (4.30) are carried

out. The average of the differences, $\overline{\epsilon}_i$, $i = RN, MD, RRT$, between simulation and iterative method, are briefly outlined below.

Case 1: Single provider ($m_n = m_d = m_r = 1$).

In the case of single RN, single MD, and one RRT, Table 4.3 shows both of the average relative error ($\overline{\delta_i}$) and the average absolute error ($\overline{\epsilon_i}$) between simulation and SRI approaches.

Table 4.3: Identical patients: Single provider case

Number of patients	2		4		6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	1.76%	0.002	5.1%	0.013	8.50%	0.031	11.65%	0.055
$i = MD$	5.22%	0.001	14.32%	0.009	19.69%	0.019	23.47%	0.030
$i = RRT$	7.85%	0.001	16.31%	0.004	21.72%	0.009	25.49%	0.014

Case 2: Multiple RNs, single MD and one RRT ($m_n > 1$, $m_d = m_r = 1$).

Table 4.4 shows the results with 2 and 4 RNs and single MD and RRT.

Case 3: Multiple RNs, MDs, and RRTs ($m_n > 1, m_d > 1, m_r = 2$).

When both MD and RRT services also have multiple providers, both the relative and absolute errors, $\overline{\delta_i}$ and $\overline{\epsilon_i}$ are presented in Table 4.5.

Next, the more general cases, non-identical patient scenarios, are investigated. Let ρ_{RN}^{sim}, ρ_i^{sim} and ρ_i^{sri}, $i = RN, MD, RRT$, denote the utilizations of RN, MD, and RRT, obtained by simulations and iterative method, respectively. The accuracy is defined the same as in Equation (4.31) and (4.32).

Case 4: Non-identical patients: single provider ($m_n = m_d = m_r = 1$).

Table 4.4: Identical patients: Multiple RNs and single MD and RRT

Number of patients	4		6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	0.55%	0.001	2.28%	0.009	5.47%	0.028
$i = MD$	0.96%	0.001	3.78%	0.003	10.15%	0.011
$i = RRT$	1.79%	0.001	7.89%	0.003	17.84%	0.008

(a) $m_n = 2$

Number of patients	6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	0.89%	0.004	9.84%	0.053
$i = MD$	3.03%	0.002	12.4%	0.013
$i = RRT$	1.01%	0.001	24.1%	0.011

(b) $m_n = 4$

Table 4.5: Identical patients: Multiple RNs, two MDs, and RRTs

Number of patients	6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	0.27%	0.001	0.19%	0.001
$i = MD$	0.87%	0.001	0.79%	0.001
$i = RRT$	1.90%	0.001	2.03%	0.001

(a) $m_n = 4, m_d = m_r = 2$

The results are shown in Table 4.6 with errors.

Table 4.6: Non-identical patients: Single provider case

Number of patients	2		4		6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	1.73%	0.002	4.06%	0.011	15.06%	0.061	24.65%	0.136
$i = MD$	7.17%	0.002	17.13%	0.010	12.54%	0.010	2.17%	0.002
$i = RRT$	7.19%	0.001	16.72%	0.004	12.97%	0.004	13.14%	0.005

Case 5: Non-identical patients: multiple RNs, single MD and one RRT ($m_n > 1$, $m_d = m_r = 1$).

From Table 4.7, it is clear that the errors are smaller.

Case 6: Non-identical patients: multiple RNs, MDs, and RRTs ($m_n > 1, m_d > 1, m_r = 2$).

Again when both MD and RRT services also have multiple providers, errors between simulation and iterative method are presented in Table 4.8.

4.2.5 System Properties

Using the results obtained above, the system-theoretic properties can be investigated. Specifically, the monotonic properties of the care delivery system can help us determine the direction of changes with respect to parameter variations. Intuitively, the RN utilization should increase if RN check rate or nurse detection rate increases. Similarly, the MD and RRT utilizations are monotonically increasing with respect to MD intervention rate and RRT initialization rate, respectively. Such intuitions can be verified in the

Table 4.7: Non-identical patients: Multiple RNs and single MD and RRT

Number of patients	4		6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	0.4%	0.001	2.17%	0.010	6.58%	0.038
$i = MD$	1.57%	0.001	4.85%	0.004	20.78%	0.017
$i = RRT$	3.18%	0.001	6.43%	0.002	7.84%	0.003

(a) $m_n = 2$

Number of patients	6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	0.74%	0.003	3.27%	0.019
$i = MD$	0.16%	0.004	23.03%	0.018
$i = RRT$	7.86%	0.002	9.88%	0.003

(b) $m_n = 4$

Table 4.8: Non-identical patients: Multiple RNs, two MDs, and RRTs

Number of patients	6		8	
Error	$\overline{\delta_i}$	$\overline{\epsilon_i}$	$\overline{\delta_i}$	$\overline{\epsilon_i}$
$i = RN$	0.50%	0.002	2.98%	0.017
$i = MD$	4.53%	0.003	20.62%	0.017
$i = RRT$	5.65%	0.001	8.28%	0.003

(a) $m_n = 4, m_d = m_r = 2$

model developed.

Let $\nearrow$ indicate monotonically increasing and $\searrow$ represent monotonically decreasing. To simplify the analysis, assume all $\lambda_{k,i} = \lambda_i$, and $\mu_{k,i} = \mu_i$, $\forall k$. In this case, all $P_{k,i} = P_i$. Then the monotonicity with respect to λ_i and μ_i can be summarized as follows:

Proposition 4.3 *Under assumptions 1)-9) with $\lambda_{k,i} = \lambda_i$, $\mu_{k,i} = \mu_i$, $\forall k$, the monotonicity properties of P_i, $i = 1, \ldots, 7$, with respect to λ_i, $i = 1, \ldots, 6$, are summarized in Table 4.9, where $\mathcal{A}$ indicate piecewise monotonicity, i.e.,*

Table 4.9: Monotonicity with respect to λ_i

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
P_1	$\searrow$	$\searrow$	$\nearrow$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
P_2	$\searrow$	$\nearrow$	$\searrow$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
P_3	$\nearrow$	$\searrow$	$\nearrow$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
P_4	$\searrow$	$\nearrow$	$\nearrow$	$\searrow$	$\mathcal{B}$	$\mathcal{C}$
P_5	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\searrow$	$\mathcal{C}$
P_6	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\searrow$
P_7	$\searrow$	$\searrow$	$\searrow$	$\searrow$	$\nearrow$	$\nearrow$

$$\mathcal{A}: \quad \begin{cases} \searrow, & \textit{if } \mathcal{M} > 0, \\ \nearrow, & \textit{otherwise}, \end{cases} \tag{4.33}$$

$$\begin{aligned} \mathcal{M} &= \mu_2(\mu_5(\lambda_6+\mu_4)+\lambda_5(\mu_5+\lambda_6)) \\ &\quad -\mu_5(\lambda_5+\mu_3)(\lambda_6+\mu_4), \end{aligned}$$

$$\mathcal{B}: \quad \begin{cases} \searrow, & \textit{if } \mathcal{N} > 0, \\ \nearrow, & \textit{otherwise}, \end{cases} \tag{4.34}$$

$$\mathcal{N} = (\mu_5+\lambda_6)\mu_3 - (\lambda_6+\lambda_4)\mu_5,$$

$$\mathcal{C}: \quad \begin{cases} \searrow, & \textit{if } \mu_4 > \mu_5, \\ \nearrow, & \textit{otherwise}. \end{cases} \tag{4.35}$$

Proof: See Appendix A. ■

Similar monotonicity can be observed with respect to μ_i as well.

Proposition 4.4 *Under assumptions 1)-9) with $\lambda_{k,i} = \lambda_i$, $\mu_{k,i} = \mu_i$, $\forall k$, the monotonicity properties of P_i, $i = 1, \ldots, 7$, with respect to μ_i, $i = 1, \ldots, 5$, are summarized in Table 4.10.*

Proof: Analogously to the proof of Proposition 4.3. ■

4.3 Conclusions

In this chapter, the patient rescue process in hospital ward is studied. An analytical model based on continuous time Markov chain has been developed. We first study the single-patient scenario and then extend to multiple patients setting. A shared resource iteration method is proposed. The iteration procedure is convergent and provides the

Table 4.10: Monotonicity with respect to μ_i

	μ_1	μ_2	μ_3	μ_4	μ_5
P_1	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
P_2	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
P_3	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$	$\nearrow$
P_4	$\nearrow$	$\searrow$	$\nearrow$	$\nearrow$	$\nearrow$
P_5	$\nearrow$	$\searrow$	$\searrow$	$\nearrow$	$\nearrow$
P_6	$\nearrow$	$\searrow$	$\searrow$	$\searrow$	$\nearrow$
P_7	$\nearrow$	$\searrow$	$\searrow$	$\searrow$	$\searrow$

unique solutions, which are aggregated steady state probabilities, reflecting performance measures related to RN, MD and RRT. The accuracy of estimation is presented through numerical study. The monotonic properties are also investigated. This provides the medical professionals a quantitative tool to study and improve floor operations and patient rescue processes.

Chapter 5

Case Study

5.1 Background

The University of Kentucky Chandler Medical Center (UKCMC) encompasses the Colleges of Dentistry, Health Sciences, Medicine, Nursing, Pharmacy, and Public Health, as well as the University of Kentucky Chandler Hospital, the Kentucky Childrens Hospital (UKCH), and the Centers of Excellence. The mission of the UKCMC is to help people of the Commonwealth and beyond to gain and retain good health through creative leadership and quality initiatives in education, research, and service. The University of Kentucky Chandler Hospital is a 473 bed tertiary care facility that serves as the major full service referral center for central and eastern Kentucky. UKCH has experienced tremendous growth over the past several years increasing from 19 000 discharges in fiscal year 2004 to more than 32 000 discharges in fiscal year 2011. A new facility of UKCH opened in the Spring of 2011 with substantially increased capacity. To address the challenges faced in the rapid response process in UKCH, an analytical model to evaluate its performance, identify the critical constraints, and propose potential improvement directions is of importance. The case study presented here is intended to contribute to this end. The rapid response process in UKCH acute care delivery is similar to the process introduced in chapter 3. With the help of acute care personnel, questionnaires have been designed and distributed to acute care personnel through rapid response team to

track the care delivery process. For each service, the patients declining information, the times of call for help, providers arrival times, decisions, etc., are recorded. A total of more than 160 samples have been collected in a two-month period. Based on the collected information, a preliminary model has been developed. Due to the limited sample size, some procedures rarely occur. Thus, we aggregate those processes (e.g., services by resident, fellow and attending doctors), and obtained a simplified process model, as shown in Figure 5.1. In this study, we focus on the decision time from nurse call to final decision making.

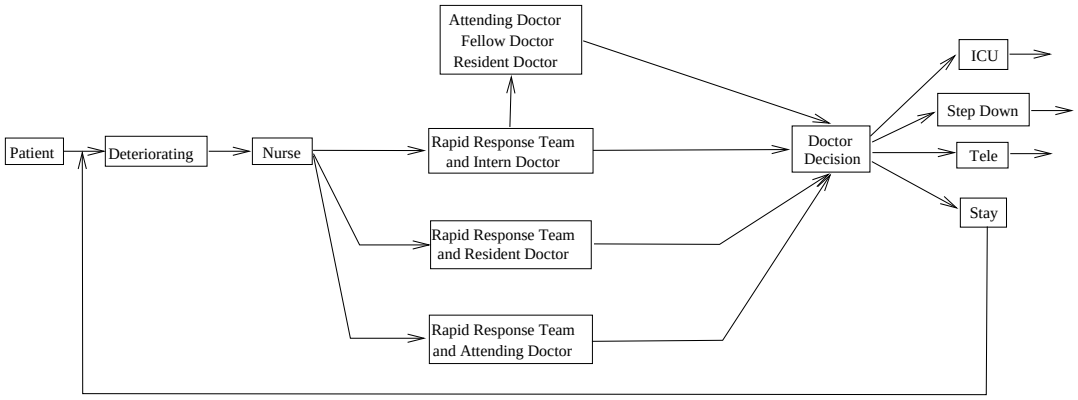

Figure 5.1: Rapid response process in acute care of UKCH

5.2 Model Validation

We first validate the model by comparing the decision time and its variabilities obtained from the model with that obtained from the collected data.

To calculate mean time and CV from Theorem 3.1 in the model, we calculate the routing probabilities α_{ij}, the average service times τ_i, and its variability cv_i. The results are shown in Tables 5.1-5.3, where afr denote the aggregated response of attending,

fellow and resident doctors.

Table 5.1: Routing probabilities

$\alpha_{r\&i,n}$	$\alpha_{r\&r,n}$	$\alpha_{r\&a,n}$	$\alpha_{afr,r\&i}$
0.30	0.50	0.15	0.29

Table 5.2: Mean response time (min)

$\tau_{r\&i}$	$\tau_{r\&r}$	$\tau_{r\&a}$	τ_{afr}
23.94	44.11	47.80	29.44

Table 5.3: CV of response time

$cv_{r\&i}$	$cv_{r\&r}$	$cv_{r\&a}$	cv_{afr}
0.92	0.68	0.91	0.78

Using these data, from (3.6) and (3.7), we calculate p_i and ρ_i as shown in Tables 5.4 and 5.5, respectively.

Then, using Theorem 3.1, the average decision time and its CV are evaluated. Let $T_d^{collect}$, $CV_d^{collect}$ denote the mean and CV obtained from the collected data, respectively, and T_d^{model}, CV_d^{model} are those calculated by Theorem 3.1. Define

$$\begin{aligned} \epsilon_T &= \frac{|T_d^{collect} - T_d^{model}|}{T_d^{collect}} \cdot 100\%, \\ \epsilon_{CV} &= \frac{|CV^{collect}(t_d) - CV_d^{model}|}{CV_d^{collect}} \cdot 100\%. \end{aligned} \tag{5.1}$$

As shown in Table 5.6, the estimate of mean decision time is accurate due to the fact that such an estimation is a weighted sum of all response times, which directly come

Table 5.4: p_i of response time

$p_{r\&i}$	$p_{r\&r}$	$p_{r\&a}$	p_{afr}
0.30	0.55	0.15	0.09

Table 5.5: ρ_i of possible routes

$\rho_{r\&i,n}$	$\rho_{r\&r,n}$	$\rho_{r\&a,n}$	$\rho_{afr,r\&i}$	$\rho_{afr,n}$
0.30	0.55	0.15	0.29	0.09

from the collected data. For CV estimation, comparing with the actual data collected on the hospital floor, the difference in CV is around 5%, which is acceptable by taking into account that there exist errors in data collection and relatively small sample size. Therefore, the model is validated.

Table 5.6: Model validation

$T_d^{collect}$	T_d^{model}	ϵ_T (%)
41.15	41.15	0

(a) Mean (min)

$CV_d^{collect}$	CV_d^{model}	ϵ_{CV} (%)
0.744	0.786	5.63

(b) CV

For the measure of response-time performance (RTP), we study the RTP as a function of desired care delivery time T_s. It is observed that the estimation of RTP is close to the data observed on the hospital floor, where the average difference is within 4%. Similar accuracy is observed when the approximate RTP is calculated based on the

aggregated response time. Thus, the model is also validated.

5.3 Improvement Analysis

Using the above model, we investigate the bottlenecks in rapid response process. From the BN-τ indicator, the RRT & resident response has the largest I_τ (where $p_{r\&r} = 0.553$), which implies this process is the BN-τ response. In addition, using the BN-cv indicator, such a response has the largest I_{cv} as well (where $p_{r\&r}\tau_{r\&r}^2 cv_{r\&r} = 0.547$). Thus, it is also the BN-cv response.

To improve the system performance, reductions of both mean and CV of RRT & Resident response are expected. Clearly, reducing the mean response time can lead to anticipated reduction of average decision time (however, such a response is still the BN-τ since it has the largest I_τ or p_i). Here, we focus on reducing variability of the response. As shown in Figure 5.2, as $cv_{r\&r}$ being reduced by 10%, CV_d decreases to 0.75, and RRT & resident is still the BN-cv. When $cv_{r\&r}$ is reduced to 58%, CV_d falls to 0.615, and BN-cv shifts to RRT & attending. Therefore, emphasizing on improving the response of RRT & resident is the key to improve the rapid response process in acute care delivery in UKCH.

We further investigate the RTP bottleneck in rapid response process by using the BN-rtp indicator. As one can see from Figure 5.3, RRT & Resident and RRT & Attending are within the BN-rtp set for all T_s considered. Since the σ is small in this study, RRT & Resident can be identified as the BN-rtp. As RRT & Resident is also the bottleneck response with respect to response time (BN-τ) and its variability (BN-cv) as discussed in Chapter 3, improving the response of RRT & Resident is the key for an efficient rapid response process in acute care delivery in UKCH. The hospital management has

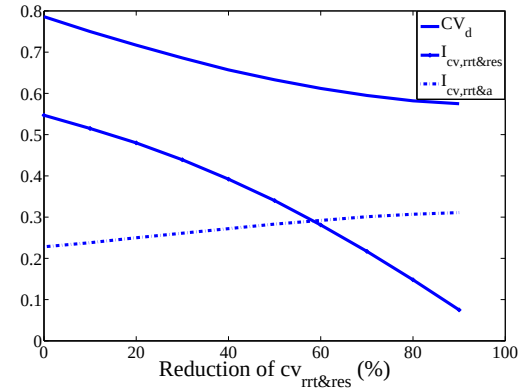

Figure 5.2: BN-*cv* in acute care of UKCH

accepted this recommendation and improvement procedures are in progress.

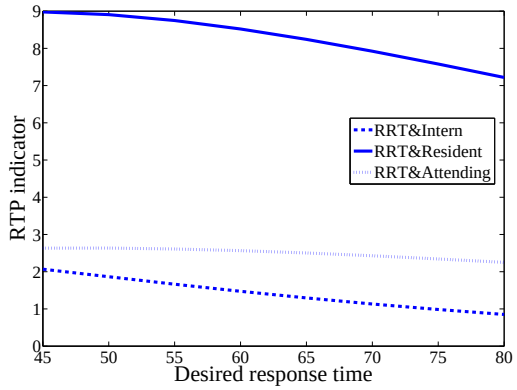

Figure 5.3: BN-*rtp* in acute care of UKCH

Chapter 6

Summary and future work

6.1 Summary

The effectiveness and efficiency of hospital rapid response system have tremendous impact on patient safety, which is always the top priority of hospital managerial decision-making. This research is the first study to address this issue from systems perspectives. Rather than using traditional clinical trial approaches or simulation modeling, analytical framework is established to quantitatively investigate two types of RRSs, which are adopted nationwide. The goal is to develop an analytical tool which can be directly applied on hospital floors for continuous improvement purposes. Detailed contributions are presented in the following two subsections.

6.1.1 RRS with Dedicated RRT

RRS with dedicated RRT features rapid response operations, which require quick delivery of care, because patient safety and care quality are strongly correlated. In light of this fact, an analytical model to characterize such operations is developed. The following three performance measures are proposed: average decision time, decision time variability and response time performance (RTP). Closed formulas are derived for performance evaluation of mean and variability for general distribution response times. For

RTP, closed formula is derived under exponential distribution. For non-exponential cases, approximation approach is provided with satisfactory accuracy. Next, monotonic properties of these three performance measures are investigated. Furthermore, the notion of bottleneck is introduced and bottleneck indicators are developed, which can help determine the most critical response under different measures directly using hospital floor data. Lastly, we address the resource sharing issue for the scenario of simultaneous declining of multiple patients. A two-level shared resource iteration method is developed. The convergence the iteration approach has been investigated analytically and numerically. The above work provides a fresh look of RRS with dedicated RRT from systems perspective and an analytical tool for hospital management.

6.1.2 RRS with Assembled RRT

Compared to RRS with dedicated RRT, RRS with assembled RRT features more correlation and coordination between multiple departments in the hospital. Therefore, resource utilization is of crucial interest. We first present a patient rescue process framework, with a five-module thinking, consisting of five modules: the triage module, patient module, floor module, RRT module, ICU module. Such five modules are interconnected and interacting with each other. We next develop a CTMC model to characterize the patient status, medical intervention by different providers, and elevation to ICU. For single patient case, closed formulas are developed to calculate the probability of the patient's staying in different states. A shared resource iteration method, is proposed to study the multiple-patient and multiple-provider scenario. Convergence property is justified numerically and accuracy is investigated. This method provides a quantitative tool to study the patient rescue process.

6.2 Future work

System modeling to study hospital rapid response is a newly developed approach. More in-depth research is needed and there exist promising research opportunities. The following problems can be studied to extend this research.

6.2.1 RRS with Dedicated RRT

The translational impact of the research is important. The bottleneck indicators for mean and CV of care delivery time can be easily calculated based on the hospital floor data, thus making the implementation attractable. However, the bottleneck analysis for response-time performance, BN-rtp, which characterizes the most critical response to impede RTP in the strongest manner, still needs to be further studied to obtain a simpler indicator, which could be directly used on hospital floor.

Another direction in this area can be focused on investigation of decision-making of the nurse, which is the first-line provider. The care quality of nurse is strongly correlated (see Ridley [91], Cronenwett et al. [31] and Aiken et al. [3]). How the experience and educational level of nurse impact the decision-making and the rapid response care quality is an issue worthy of further investigation.

In addition, it would be interesting to model patient declining during rapid response process. How the declining affect the decision-making of different responses and the care quality can be investigated. The main challenge is to model patient declining, since a number of patient's vitals jointly reflect deterioration. To the best of our knowledge, no satisfactory quantitative model exist in the literature. Therefore, it is needed for medical researchers and industrial engineering researchers to collaboratively develop such a model. By incorporating such a model, many issues in rapid response can be

further investigated.

6.2.2 RRS with Assembled RRT

First, the accuracy of the proposed iterative method can be improved. Further investigation of updating procedure and convergence property are needed.

The interactions of the modules in the analytical framework need to be further investigated. Either analytical or simulation models need to be developed to uncover the underlying principles in the system to provide improvement recommendations.

Specifically, the following work can be carried out:

- Exponential assumption is introduced in the current model. In reality, it may not hold. Thus, extending the study to non-exponential cases is of importance.

- Using the model developed, the issues of staffing level, RN checking frequency, RRT team composition, RRT call initiation, etc., can be studied. Specifically, the following questions can be raised: For a n-patient hospital ward setting, how many RNs, MDs and RRTs should be assigned? How does patient status affect the checking frequency of RNs? What is the RRT calling criteria that can effectively improve patient safety and operation efficiency of the providers simultaneously? Answers to these questions can help build a more effective and efficient hospital rapid response system.

- Again, modeling patient status and his or her declining process is important. The current model only describes two states, risk and non-risk. More detailed continuous modeling of declining process can provide useful information for provider interventions.

- Such a model can help investigate coordinations between different hospital departments or units, and different providers, etc,.
- Moreover, such a study will lay a solid foundation to study patient care delivery and patient flow in a whole hospital setting.

Finally, as rapid response operations involve strong time dynamics, transient analysis is needed, in particular, when patient declining is included in the model. However, transient behavior is less studied not only in rapid response systems, but also in almost all engineering and service systems due to its extreme difficulty. Development in this area will not only provide methods and tools for hospital rapid response and patient rescue, but also contribute substantially to all areas, which will open up a new research field.

In summary, the study of the aforementioned work will further form a more complete and powerful analytical tool for performance analysis and continuous improvement of rapid response operations, which is of crucial significance due to its strong correlation with patient safety.

Appendix

Appendix A: Proofs

Proofs of Chapter 3

Proof of Theorem 3.1: The mean of decision time follows immediately by summing up all possible responses with weights.

The variance of decision time can be evaluated as follows:

$$\begin{aligned} Var(t_d) &= Var\left(\sum_{i\in X} A_i t_i\right) \\ &= \sum_{i\in X} Var(A_i t_i) + \sum_{i\in X}\sum_{j\in X, j\neq i} Cov(A_i t_i, A_j t_j). \end{aligned} \tag{A.1}$$

The first term in (A.1) can be rewritten as:

$$\begin{aligned} Var(A_i t_i) &= E(A_i^2 t_i^2) - E^2(A_i t_i) \\ &= E(A_i^2)E(t_i^2) - E^2(A_i)\tau_i^2 \\ &= E(A_i^2)[Var(t_i) + \tau_i^2] - E^2(A_i)\tau_i^2. \end{aligned}$$

Since only one route will be selected, A_i can only take value of 1 or 0. Then,

$$\begin{aligned} E(A_i) &= 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i, \\ E(A_i^2) &= 1^2 \cdot p_i + 0^2 \cdot (1 - p_i) = p_i. \end{aligned}$$

it follows that

$$\begin{aligned} Var(A_i t_i) &= p_i(Var(t_i) + \tau_i^2) - p_i^2\tau_i^2 \\ &= p_i[Var(t_i) + (1 - p_i)\tau_i^2]. \end{aligned} \tag{A.2}$$

The second term in (A.1) can be rewritten as:

$$\begin{aligned}\sum_{j\in X, j\neq i} Cov(A_i t_i, A_i t_j) &= \sum_{j\in X, j\neq i} [E(A_i t_i \cdot A_j t_j) - E(A_i t_i)E(A_j t_j)] \\ &= \sum_{j\in X, j\neq i} [E(A_i A_j)\tau_i\tau_j - E(A_i)E(A_j)\tau_i\tau_j]\end{aligned}$$

If $j = n$, then $A_j = 1$ with probability 1 since nurse response is always the first one. Thus,

$$E(A_i A_j) = E(A_i)E(A_j) = E(A_i).$$

It follows that

$$\sum_{j\in X, j\neq i} Cov(A_i t_i, A_i t_j) = 0.$$

If $j \neq n$, again since there is only one patient in the system at any time and only one route will be selected, thus, $A_i A_j = 1$ if the patient go through the route between service j to service i, with probability ρ_{ij}. Then we have

$$E(A_i A_j) = p_j \cdot \rho_{ij} = p_j \rho_{ij}.$$

Finally, we obtain

$$\sum_{j\in X, j\neq i} Cov(A_i t_i, A_i t_j) = \sum_{j\in X, j\neq i,n} (\rho_{ij} p_j \tau_i \tau_j - p_i p_j \tau_i \tau_j). \tag{A.3}$$

By combining (A.2) and (A.3) and replacing them in (A.1), we have

$$\begin{aligned}Var(t_d) &= \sum_{i\in X} p_i[Var(t_i) + (1-p_i)\tau_i^2] \\ &\quad + \sum_{i\in X} \sum_{j\in X, j\neq i,n} (\rho_{ij} p_j \tau_i \tau_j - p_i p_j \tau_i \tau_j) \\ &= \sum_{i\in X} p_i \tau_i^2 (cv_i^2 + 1 - p_i) \\ &\quad - \sum_{i\in X} \tau_i \Big[\sum_{j\in X, j\neq i,n} (p_i - \rho_{ij}) p_j \tau_j \Big].\end{aligned}$$

It follows that

$$\begin{aligned} CV_d &= \frac{\sqrt{Var(t_d)}}{T_d} \\ &= \frac{1}{T_d}\Big(\sum_{i\in X} p_i\tau_i^2(cv_i^2+1-p_i) \\ &\quad -\sum_{i\in X}\tau_i\Big[\sum_{j\in X, j\neq i,n}(p_i-\rho_{ij})p_j\tau_j\Big]\Big)^{\frac{1}{2}}. \end{aligned}$$

■

Proof of Theorem 3.2: Assuming there are K decision routes, where routes $1,\ldots,k_1$ are direct routes (i.e., no higher level help is requested), and $k_1+1,\ldots,K$ are the routes which require further help. Denote l_i, $i=1,\ldots,K$, as the routing indicator. Then $l_i=1$ indicates that route i is selected, and $l_i=0$ not selected. Let t_d denote the decision time, and $t_d(i)$ represent the decision time when route i is selected. Then, RTP can be expressed as follows:

$$\begin{aligned} Pr(t_d\le T_s) &= \sum_{i=1}^{K} Pr(t_d\le T_s|l_i=1)Pr(l_i=1) \\ &= \sum_{i=1}^{K} Pr(t_d(i)\le T_s)Pr(l_i=1) \\ &= \sum_{i=1}^{k_1} Pr(t_d(i)\le T_s)Pr(l_i=1) \\ &\quad + \sum_{i=k_1+1}^{K} Pr(t_d(i)\le T_s)Pr(l_i=1). \end{aligned} \tag{A.4}$$

For the first term, when $i\le k_1$, only one response (assuming j) is initiated. Therefore,

$$Pr(t_d(i)\le T_s) = 1-e^{-\frac{T_s}{\tau_j}}, \quad i=1,\ldots,k_1,$$

and

$$Pr(l_i=1)=\alpha_{i,n}. \tag{A.5}$$

For the second term, it involves more than one responses. Thus, the total response time follows a hypoexponential distribution. For a route i, $i = k_1 + 1, \ldots, K$, with m responses in total, assume it starts from a nurse call to response j and finishes with response q. The intermediate responses are denoted as $r_1, r_2, \ldots, r_{m-2}$. Then, we have

$$\begin{aligned} Pr(t_d(i) \leq T_s) &= 1 - C_j e^{-\frac{T_s}{\tau_j}} - C_q e^{-\frac{T_s}{\tau_q}} \\ &\quad - \sum_{s=1}^{m-2} C_{r_s} e^{-\frac{T_s}{\tau_{r_s}}}, \quad i = k_1 + 1, \ldots, K, \end{aligned}$$

where

$$C_i = \sum_{i \in X} \prod_{j \in X, j \neq i} \frac{\tau_j}{\tau_i - \tau_j}.$$

In addition,

$$Pr(l_i = 1) = \rho_{qj} \alpha_{j,n}, \tag{A.6}$$

where ρ_{qj} is the probability that a patient goes through services j to q, which is defined in Equation (3.7).

Replacing the above expressions back to Equation (A.4), the RTP calculation formula is obtained. ■

Proof of Proposition 3.1:

Immediately obtained from Theorem 3.1 by showing that

$$\frac{\partial E(t_d)}{\partial \tau_i} = p_i > 0. \tag{A.7}$$

■

Proof of Proposition 3.2: From

$$\frac{\partial CV_d}{\partial cv_i} = \frac{\partial \left(\frac{\sqrt{Var(t_d)}}{T_d} \right)}{\partial \left(\frac{\sqrt{Var(t_i)}}{\tau_i} \right)},$$

since τ_i is unchanged, by Theorem 3.1, T_d will still be the same. Thus, we have

$$\begin{aligned}
\frac{\partial CV_d}{\partial cv_i} &= \frac{\tau_i}{T_d} \cdot \frac{\partial \sqrt{Var(t_d)}}{\partial \sqrt{Var(t_i)}} = \frac{\tau_i}{T_d} \cdot \frac{\frac{\partial \sqrt{Var(t_d)}}{\partial Var(t_i)}}{\frac{\partial \sqrt{Var(t_i)}}{\partial Var(t_i)}} \\
&= \frac{\tau_i}{T_d} \cdot \frac{\frac{\partial \sqrt{Var(t_d)}}{\partial Var(t_d)} \cdot \frac{\partial Var(t_d)}{\partial Var(t_i)}}{\frac{1}{2\sqrt{Var(t_i)}}} \\
&= \frac{\tau_i}{T_d} \cdot \frac{\frac{1}{2\sqrt{Var(t_d)}} \cdot \frac{\partial Var(t_d)}{\partial Var(t_i)}}{\frac{1}{2\sqrt{Var(t_i)}}} \\
&= \frac{\tau_i}{T_d} \cdot \frac{\sqrt{Var(t_i)}}{\sqrt{Var(t_d)}} \cdot \frac{\partial Var(t_d)}{\partial Var(t_i)}.
\end{aligned}$$

From (3.9) in Theorem 3.1,

$$\frac{\partial Var(t_d)}{\partial Var(t_i)} = p_i.$$

Thus, we obtain

$$\begin{aligned}
\frac{\partial CV_d}{\partial cv_i} &= p_i \cdot \frac{\tau_i}{T_d} \cdot \frac{\sqrt{Var(t_i)}}{Var(t_d)} \\
&= p_i \cdot \frac{cv_i}{CV_d} \cdot \frac{\tau_i^2}{T_d^2} \qquad \text{(A.8)} \\
&> 0.
\end{aligned}$$

■

Proof of Proposition 3.4: From (A.7), it is clearly that $\frac{\partial T_d}{\partial \tau_i}$ is maximized if and only if p_i is maximized. ■

Proof of Proposition 3.5: From (A.9), it is easy to see that $\frac{\partial CV_d}{\partial cv_i}$ is maximized if and only if $p_i \cdot \frac{cv_i}{CV_d} \cdot \frac{\tau_i^2}{T_d^2}$ is maximized. By deleting the common terms T_d and CV_d, the argument is arrived. ■

Proof of Proposition 3.6: Under condition (3.21), RTP can be expressed as follows:

$$RTP = \sum_{i \in X, i \neq f,a} \left(1 - e^{-\frac{T_s}{\tau_i}}\right) \alpha_{i,n} \tag{A.9}$$

Then, taking the partial derivatives of τ_i, we obtain the following:

$$\left|\frac{\partial RTP}{\partial \tau_i}\right| = \left|\alpha_{i,n} \cdot (e^{-\frac{T_s}{\tau_i}}) \cdot T_s \cdot (\frac{1}{\tau_i^2})\right|.$$

Similar derivation is applied to τ_j. Thus, ignoring the common term T_s, the comparison formula is obtained. ■

To prove Theorem 3.3, Lemma A.1 and Lemma A.2 are needed.

Lemma A.1 *If* $p_{2,r}^{(j)} > p_{2,r}^{(j-1)}$, $i = 1, 2$; $r \in X_1$; $j = 1, 2, 3, ...,$ *we have* $\tau_{1,r}^{(j+1)} > \tau_{1,r}^{(j)}$, $p_{1,r}^{(j+1)} < p_{1,r}^{(j)}$, $\tau_{2,r}^{(j+1)} < \tau_{2,r}^{(j)}$, $p_{2,r}^{(j+1)} > p_{2,r}^{(j)}$.

Proof of Lemma A.1: From all the equations related to the update of $\tau_{i,r}^{(j)}$ and $p_{i,r}^{(j)}$, which are from (3.71) to (3.99), we have the following equations, with $C_{1,r}$ and $C_{2,r}$ being constants related to resource r, respectively. Specifically, we have

$$C_{1,r} = \begin{cases} p_{int}\tau_{int} + p_{rrt\&int}\tau_{rrt\&int}, & \text{if } r = int, \\ p_{res}\tau_{res} + p_{rrt\&res}\tau_{rrt\&res} & \text{if } r = res, \\ p_{rrt}\tau_{rrt} + p_{rrt\&int}\tau_{rrt\&int} + p_{rrt\&res}\tau_{rrt\&res} & \\ +p_{rrt\&f}\tau_{rrt\&f} + p_{rrt\&a}\tau_{rrt\&a} & \text{if } r = rrt, \\ p_f\tau_f + p_{rrt\&f}\tau_{rrt\&f} & \text{if } r = f, \\ p_a\tau_a + p_{rrt\&a}\tau_{rrt\&a} & \text{if } r = a. \end{cases}$$

$$C_{2,r} = \begin{cases} p_{int}^2\tau_{int} + p_{rrt\&int}^2\tau_{rrt\&int}, & \text{if } r = int, \\ p_{res}^2\tau_{res} + p_{rrt\&res}^2\tau_{rrt\&res} & \text{if } r = res, \\ p_{rrt}^2\tau_{rrt} + p_{rrt\&int}^2\tau_{rrt\&int}^2 + p_{rrt\&res}^2\tau_{rrt\&res} & \\ +p_{rrt\&f}^2\tau_{rrt\&f} + p_{rrt\&a}^2\tau_{rrt\&a} & \text{if } r = rrt, \\ p_f^2\tau_f + p_{rrt\&f}^2\tau_{rrt\&f} & \text{if } r = f, \\ p_a^2\tau_a + p_{rrt\&a}^2\tau_{rrt\&a} & \text{if } r = a. \end{cases}$$

For iteration j, if $p_{2,r}^{(j)} > p_{2,r}^{(j-1)}$, then for patient 1:

$$\tau_{1,r}^{(j)} = T_d + p_{2,r}^{(j-1)}C_{1,r} < T_d + p_{2,r}^{(j)}C_{1,r} = \tau_{1,r}^{(j+1)}, \tag{A.10}$$

$$p_{1,r}^{(j)} = \frac{C_{1,r}}{\tau_{1,r}^{(j)}} > \frac{C_{1,r}}{\tau_{1,r}^{(j+1)}} = p_{1,r}^{(j+1)}. \tag{A.11}$$

This leads to, for patient 2,

$$\tau_{2,r}^{(j)} = T_d + p_{1,r}^{(j)}C_{1,r} > T_d + p_{1,r}^{(j+1)}C_{1,r} = \tau_{2,r}^{(j+1)}, \tag{A.12}$$

$$p_{2,r}^{(j)} = \frac{C_{2,r}}{\tau_{2,r}^{(j)}} < \frac{C_{2,r}}{\tau_{2,r}^{(j+1)}} = p_{2,r}^{(j+1)}. \tag{A.13}$$

The obtained results in the above four inequations complete the proof. ■

Lemma A.2 *For $i = 1, 2$; $r \in X_1$; $j = 1, 2, 3, ...$, the sequences $p_{1,r}^{(j)}$ and $\tau_{2,r}^{(j)}$ are monotonically decreasing, while the sequences $p_{2,r}^{(j)}$ and $\tau_{1,r}^{(j)}$ are monotonically increasing.*

Proof of Lemma A.2: Mathematical induction is used to proof this lemma. Initial Step: When $j = 1$, since $p_{2,r}^{(0)} = 0$, from Equation(A.13), we have

$$p_{2,r}^{(1)} > p_{2,r}^{(0)} = 0, \tag{A.14}$$

Then, from Lemma A.1, we have

$$\tau_{1,r}^{(2)} > \tau_{1,r}^{(1)}, \tag{A.15}$$

$$p_{1,r}^{(2)} < p_{1,r}^{(1)}, \tag{A.16}$$

$$\tau_{2,r}^{2)} < \tau_{2,r}^{(1)}, \tag{A.17}$$

$$p_{2,r}^{(2)} > p_{2,r}^{(1)}. \tag{A.18}$$

The above proves the base case. Inductive Step: Assume when $j = k$,

$$\tau_{1,r}^{(k+1)} > \tau_{1,r}^{(k)}, \tag{A.19}$$

$$p_{1,r}^{(k+1)} < p_{1,r}^{(k)}, \tag{A.20}$$

$$\tau_{2,r}^{(k+1)} < \tau_{2,r}^{(k)}, \tag{A.21}$$

$$p_{2,r}^{(k+1)} > p_{2,r}^{(k)}. \tag{A.22}$$

From Lemma A.1, this leads to

$$\tau_{1,r}^{(k+2)} > \tau_{1,r}^{(k+1)}, \tag{A.23}$$

$$p_{1,r}^{(k+2)} < p_{1,r}^{(k+1)}, \tag{A.24}$$

$$\tau_{2,r}^{(k+2)} < \tau_{2,r}^{(k+1)}, \tag{A.25}$$

$$p_{2,r}^{(k+2)} > p_{2,r}^{(k+1)}. \tag{A.26}$$

Thus, the case of $j = k + 1$ also holds. Therefore, by induction, we obtain that, for $i = 1, 2$; $r \in X_1$; $j = 1, 2, 3, ...$, the sequences $p_{1,r}^{(j)}$ and $\tau_{2,r}^{(j)}$ are monotonically decreasing, while the sequences $p_{2,int}^{(j)}$ and $\tau_{1,int}^{(j)}$ are monotonically increasing. ■

Proof of Theorem 3.3: From Lemma A.2, we obtain that for $i = 1, 2$; $r \in X_1$; $j = 1, 2, 3, ...$, the sequences $p_{1,r}^{(j)}$ and $\tau_{2,r}^{(j)}$ are monotonically decreasing, while the sequences

$p_{2,int}^{(j)}$ and $\tau_{1,int}^{(j)}$ are monotonically increasing. Next we show that the sequences $\tau_{i,r}^{(j)}$ and $p_{i,r}^{(j)}$, $i = 1, 2$; $r \in X_1$; $j = 1, 2, 3, ...$, are bounded from above and below. For $p_{i,r}^{(j)}$s, from Equations (A.11) and (A.13), we have

$$0 < p_{i,r}^{(j)} < 1. \tag{A.27}$$

For $\tau_{i,r}^{(j)}$s, from Equations (A.10) and (A.12), since $0 < p_{i,r}^{(j)} < 1$, we obtain

$$T_d < \tau_{i,r}^{(j)} < T_d + C_{i,r}. \tag{A.28}$$

Since the sequences $\tau_{i,r}^{(j)}$ and $p_{i,r}^{(j)}$, $i = 1, 2$; $r \in X_1$; $j = 1, 2, 3, ...$ are monotonic and bounded from above and below, they are convergent, which proves Equations (3.116) and (3.117) hold in the level-1 iteration. ■

To prove Theorem 3.4, Lemma A.3 and Lemma A.4 are needed.

Lemma A.3 *$g_i^{(l)} > g_i^{(l-1)}$, $i = 1, ..., n$, $l = 1, 2, ...$, if and only if $\mu_i^{(l)} > \mu_i^{(l-1)}$.*

Proof of Lemma A.3: From Equations (3.107), (3.109) and (3.111), for $i = 1, ..., n$, $l = 1, 2, ...$, we have

$$g_i^{(l)} = \frac{\mu_i^{(l)}}{\mu_i^{(l)} + T_{normal}}. \tag{A.29}$$

Equation (A.29) can be rewritten as

$$\mu_i^{(l)} = \frac{T_{normal}}{\frac{1}{g_i^{(l)}} - 1} > \frac{T_{normal}}{\frac{1}{g_i^{(l-1)}} - 1} = \mu_i^{(l-1)}. \tag{A.30}$$

To prove the second argument, Equation (A.29) can be rewritten as

$$g_i^{(l)} = \frac{1}{1 + T_{normal}/\mu_i^{(l)}} > \frac{1}{1 + T_{normal}/\mu_i^{(l-1)}} = g_i^{(l-1)}. \tag{A.31}$$

Therefore, Lemma A.3 holds. ■

Lemma A.4 *If* $g_i^{(l)} > g_i^{(l-1)}$, $i = 1, ..., n$, $l = 1, 2, ...,$ *then* $g_i^{(l+1)} > g_i^{(l)}$.

Proof of Lemma A.4: We first prove that when $i = 1$, $l = 1, 2, ...,$

$$g_1^{(l+1)} > g_1^{(l)}. \tag{A.32}$$

From Equation (3.106), we obtain

$$\mu_1^{(l+1)} = T_{in}(1 + g_1^{(l)}\Sigma_{i=2,...,n}g_i^{(l)}) > T_{in}(1 + g_1^{(l-1)}\Sigma_{i=2,...,n}g_i^{(l-1)}) = \mu_1^{(l)} \tag{A.33}$$

Therefore, from Lemma A.3, inequation (A.32) holds. When $2 \leq i \leq n-1$, we prove by contradiction. Suppose the statement in Lemma A.4 is false, i.e., for $i = 2, ..., n-1$, $l = 1, 2, ...,$ if $g_i^{(l)} > g_i^{(l-1)}$, there exists an i^*, $2 \leq i^* \leq n-1$, that makes the following inequation hold:

$$g_{i^*}^{(l+1)} \leq g_{i^*}^{(l)}, \tag{A.34}$$

and for all $i > i^*$,

$$g_i^{(l+1)} > g_i^{(l)}, \tag{A.35}$$

Then from Lemma A.3, we have

$$\mu_{i^*}^{(l+1)} \leq \mu_{i^*}^{(l)}. \tag{A.36}$$

From Equation (3.108) and (3.109), it follows that

$$\mu_{i^*}^{(l+1)} = T_{in}(1 + g_{i^*}^{(l)}(\Sigma_{j=1,...,i^*-1}g_j^{(l+1)} + \Sigma_{j=i^*+1,...,n}g_j^{(l)})), \tag{A.37}$$

$$\mu_{i^*}^{(l)} = T_{in}(1 + g_{i^*}^{(l-1)}(\Sigma_{j=1,...,i^*-1}g_j^{(l)} + \Sigma_{j=i^*+1,...,n}g_j^{(l-1)})). \tag{A.38}$$

Since $g_i^{(l)} > g_i^{(l-1)}$, for any i and l, the following two inequations hold.

$$g_{i^*}^{(l)} > g_{i^*}^{(l-1)}, \tag{A.39}$$

$$\Sigma_{j=i^*+1,...,n}g_j^{(l)} > \Sigma_{j=i^*+1,...,n}g_j^{(l-1)}. \tag{A.40}$$

Therefore, we must have

$$\Sigma_{j=1,\ldots,i^*-1} g_j^{(l+1)} > \Sigma_{j=1,\ldots,i^*-1} g_j^{(l)} \tag{A.41}$$

From Equations (A.37) and (A.38), this implies that, there should exists at least one $i^{**} < i^*$, such that the following inequation holds:

$$g_{i^{**}}^{(l+1)} \leq g_{i^{**}}^{(l)}, \tag{A.42}$$

And for all $i > i^{**}$,

$$g_i^{(l+1)} > g_i^{(l)}. \tag{A.43}$$

Furthermore from Lemma A.3, we have

$$\mu_{i^{**}}^{(l+1)} \leq \mu_{i^{**}}^{(l)}. \tag{A.44}$$

Applying the same logic again, if Equation (A.44) holds, we can find the largest $i^{***} < i^{**}$ making the following inequation hold.

$$g_{i^{***}}^{(l+1)} \leq g_{i^{***}}^{(l)}, \tag{A.45}$$

and

$$\mu_{i^{***}}^{(l+1)} \leq \mu_{i^{***}}^{(l)}. \tag{A.46}$$

By introduction, this follows that for any given $2 \leq i^* \leq n-1$, with $g_{i^*}^{(}l+1) \leq g_{i^*}^{(}l)$. there always exists an i^{**} such that

$$g_{i^{**}}^{(l+1)} \leq g_{i^{**}}^{(l)}, \tag{A.47}$$

$$\mu_{i^{**}}^{(l+1)} \leq \mu_{i^{**}}^{(l)}. \tag{A.48}$$

This implies that

$$g_2^{(l+1)} \leq g_2^{(l)}, \tag{A.49}$$

$$\mu_2^{(l+1)} \leq \mu_2^{(l)}. \tag{A.50}$$

However, from Equations (3.108) and (3.109), we have

$$\mu_2^{(l+1)} = T_{in}(1 + g_2^{(l)}(g_1^{(l+1)} + \Sigma_{j=3,...,n} g_j^{(l)})), \tag{A.51}$$

$$\mu_2^{(l)} = T_{in}(1 + g_2^{(l-1)}((g_1^{(l)} + \Sigma_{j=3,...,n} g_j^{(l-1)})). \tag{A.52}$$

From inequation (A.32), we obtain

$$\mu_2^{(l+1)} > \mu_2^{(l)}, \tag{A.53}$$

which contradicts to $\mu_2^{(l+1)} \leq \mu_2^{(l)}$. Therefore, by contradiction, we can show that when $2 \leq i \leq n-1$, $l = 1, 2, ...$, $g_l^{(l+1)} > g_l^{(l)}$.

By induction, we can show that for $i = 2, ..., n$, $g_i^{(l+1)} > g_i^{(l)}$. For $i = 1, ..., n$, $l = 1, 2, ...$, if $g_i^{(l)} > g_i^{(l-1)}$ holds, we have $g_i^{(l+1)} > g_i^{(l)}$. Lemma A.4 is proved. ■

Proof of Theorem 3.4: First we prove that the sequences $\mu_i^{(l)}$ and $g_i^{(l)}$, $i = 1, 2, ...n$; $l = 1, 2, 3, ...$, are monotonically increasing using mathematical induction. Initial Step: When $l = 1$, since $g_i^{(0)} = 0$, from Lemma A.4, $g_i^{(1)} > g_i^{(0)} = 0$, then $g_i^{(2)} > g_i^{(1)}$ and $\mu_i^{(2)} > \mu_i^{(1)}$. The base case is proved. Inductive Step: The inductive assumption is that when $l = k$, we have $\mu_i^{(l)} > \mu_i^{(l-1)}$ and $g_i^{(l)} > g_i^{(l-1)}$, $i = 1, 2, ...n$; $l = 1, 2, , ..., k$. Then from Lemma A.4 and using the condition $g_i^{(k)} > g_i^{(k-1)}$, we have $\mu_i^{(k+1)} > \mu_i^{(k)}$ and $g_i^{(k+1)} > g_i^{(k)}$. Therefore, the case where $l = k + 1$ also holds. Therefore, the sequences $\mu_i^{(l)}$ and $g_i^{(l)}$, $i = 1, 2, ...n$; $l = 1, 2, 3, ...$, are monotonically increasing. For boundedness, $g_i^{(l)}$s are bounded between 0 and 1 from Equation (3.107), (3.109), and (3.111), while $\mu_i^{(l)}$s are also bounded according to Equations (3.106) and (3.108). Since the sequences $\mu_i^{(l)}$ and $g_i^{(l)}$, $i = 1, 2, ..., n$, are both monotonic and bounded from above and below, they are convergent, which proves Equations (3.118) and (3.119). ■

Proofs of Chapter 4

To prove Proposition 4.1, it is convenient to present the proof of Proposition 4.3 first below.

Proof of Proposition 4.3:

Under assumptions 1)-9), we investigate with relationship of P_i, $i = 1, \ldots, 7$, with respect to λ_k, $k = 1, \ldots, 6$. Due to space limitation, we provide here only the proof of P_1. The proofs of other $_i$s can be obtained similarly. The expression for P_1 is shown as follows:

$$P_1 = \frac{1}{\Psi}$$

where,

$$\begin{aligned}
\Psi &= C_1 + C_2 + C_3 + C_2C_4 + C_2C_4C_5 \\
&\quad + C_2C_4C_5C_6 + C_2C_4C_5C_6C_7, \\
C_1 &= 1, \\
C_2 &= \lambda_2/\lambda_3, \\
C_3 &= \lambda_1/\mu_1, \\
C_4 &= \lambda_3/(\lambda_4 + \mu_2), \\
C_5 &= \lambda_4/(\lambda_5 + \mu_3), \\
C_6 &= \lambda_5/(\lambda_6 + \mu_4), \\
C_7 &= \lambda_6/\mu_5.
\end{aligned}$$

First, consider λ_1. Rewrite P_1 as

$$P_1 = \frac{1}{A_1\lambda_1 + B_1},$$

where A_1 and B_1 are terms without λ_1. Specifically,

$$\begin{aligned} A_1 &= \frac{1}{\mu_1}, \\ B_1 &= C_1 + C_2 + C_4 + C_5 + C_6 + C_7. \end{aligned}$$

Then, taking partial derivative of P_1 with respect to λ_1, we obtain

$$\frac{\partial P_1}{\partial \lambda_1} = -\frac{A}{(A\lambda_1 + B)^2} < 0.$$

Therefore, P_1 is monotonically decreasing with respect to λ_1.

Next, consider λ_2. Using the exact same approach, it is easy to show that P_1 is monotonically decreasing with respect to λ_2.

Then, consider λ_3. Rewrite P_1 as

$$P_1 = \frac{1}{\frac{A_3}{\lambda_3} + B_3} < 0,$$

where A_3 and B_3 are terms without λ_3. Specifically,

$$\begin{aligned} A_3 &= \lambda_2, \\ B_3 &= C_1 + C_3 + C_4 + C_5 + C_6 + C_7. \end{aligned}$$

Then, by taking partial derivative of P_1 with respect to λ_3, we have

$$\frac{\partial P_1}{\partial \lambda_3} = \frac{(-1)}{(A_3\lambda_3 + B_3)^2} A_3 \frac{(-1)}{\lambda_3^2} > 0.$$

Therefore, P_1 is monotonically increasing with respect to λ_3.

Now we consider λ_4. Again rewrite P_1 as:

$$P_1 = \frac{1}{A_4 + \frac{\lambda_2}{\lambda_4+\mu_2} + \frac{\lambda_2\lambda_4}{(\lambda_4+\mu_2)B_4} + \frac{\lambda_2\lambda_4 E_4}{(\lambda_4+\mu_2)B_4} + \frac{\lambda_2\lambda_4 E_4 D_4}{(\lambda_4+\mu_2)B_4}},$$

where A_4, B_4, D_4 and E_4 are terms without λ_4. In particular,

$$\begin{aligned} A_4 &= C_1 + C_2 + C_3, \\ B_4 &= \lambda_5 + \mu_3, \\ D_4 &= \frac{\lambda_6}{\mu_5}, \\ E_4 &= \frac{\lambda_5}{\lambda_6 + \mu_4}. \end{aligned}$$

After further simplification of P_1, we have

$$P_1 = \frac{1}{A_4 + \frac{F_4\lambda_4 + B_4\lambda_2}{B_4(\lambda_4+\mu_2)}},$$

and

$$F_4 = \lambda_2 + E_4\lambda_2 + E_4D_4\lambda_2.$$

Taking partial derivative of P_1 with respect to λ_4, it leads to

$$\frac{\partial P_1}{\partial \lambda_4} = \frac{(-1)}{(A_4 + \frac{F_4\lambda_4 + B_4\lambda_2}{B_4^2(\lambda_4+\mu_2)^2}} \cdot \frac{F_4\mu_2 - B_4\lambda_2}{B_4(\lambda_4+\mu_2)^2}.$$

Determining the sign of this partial derivative is equivalent to determine the sign of

$$\Psi = F_4\mu_2 - B_4\lambda_2.$$

By substituting the expressions of B_4, D_4, E_4 and F_4, it follows that

$$\begin{aligned} \Psi &= \lambda_2(1 + E_4 + E_4D_4)\mu_2 - \lambda_2(\lambda_5 + \mu_3) \\ &= \lambda_2[(1 + C + E_4D_4)\mu_2 - (\lambda_5 + \mu_3)] \\ &= \lambda_2[(1 + \frac{\lambda_5}{\lambda_6 + \mu_4} + \frac{\lambda_5}{(\lambda_6 + \mu_4)}\frac{\lambda_6}{\mu_5})\mu_2 - (\lambda_5 + \mu_3)] \\ &= \frac{\lambda_2[(\lambda_6 + \mu_4)\mu_5 + \lambda_5(\mu_5 + \lambda_6)]\mu_2}{(\lambda_6 + \mu_4)\mu_5} \\ &\quad - \frac{\lambda_2(\lambda_5 + \mu_3)\mu_5(\lambda_6 + \mu_4)}{(\lambda_6 + \mu_4)\mu_5} \\ &= \frac{1}{(\lambda_6 + \mu_4)\mu_5}M, \end{aligned}$$

where

$$M = \mu_2[(\lambda_6 + \mu_4)\mu_5 + \lambda_5(\mu_5 + \lambda_6)] - \mu_5(\lambda_5 + \mu_3)(\lambda_6 + \mu_4).$$

Therefore, when $M > 0$, then $\Psi > 0$, and $\frac{\partial P_1}{\partial \lambda_4} < 0$, which implies that P_1 is monotonically decreasing with respect to λ_4. When $M < 0$, P_1 is monotonically increasing with respect to λ_4.

Similarly, the proofs the monotonicity property of P_1 with respect to λ_5 and λ_6 can be carried out. ■

Lemma A.5 is needed to prove Proposition 4.1.

Lemma A.5 *For $i = 1, 2$; $j = 1, 2, 3, ...$, the sequences $\lambda_{1,4}^{(j)}$ and $P_{1,5}^{(j)}$ are monotonically increasing, while the sequences $\lambda_{2,4}^{(j)}$ and $P_{2,5}^{(j)}$ are monotonically decreasing.*

Proof of Lemma A.5: Under assumptions 1)-9), the updating equations related to MD during iteration j can be written as follows:

For patient 1:

$$\lambda_{1,4}^{(j+1)} = \lambda_{1,4}^{(0)}\Big(1 - P_{2,5}^{(j)}\Big), \tag{A.54}$$

then, $P_{1,5}^{(j+1)}$ can be obtained as follows.

$$P_{1,5}^{(j+1)} \;=\; f_{P,5}\Big(\Lambda_1^{(j+1)}, \Gamma_1\Big), \tag{A.55}$$

where $\lambda_{1,4}^{(j+1)}$ is used to replace $\lambda_{1,4}^{(j)}$ in $\Lambda_1^{(j+1)}$.

For patient 2:

$$\lambda_{2,4}^{(j+1)} = \lambda_{2,4}^{(0)}\Big(1 - P_{1,5}^{(j+1)}\Big), \tag{A.56}$$

then, $P_{2,5}^{(j+1)}$ can be obtained as follows.

$$P_{2,5}^{(j+1)} \;=\; f_{P,5}\Big(\Lambda_2^{(j+1)}, \Gamma_2\Big), \tag{A.57}$$

where $\lambda_{2,4}^{(j+1)}$ is used to replace $\lambda_{2,4}^{(j)}$ in $\Lambda_2^{(j+1)}$.

Mathematical induction is then used. Initial Step ($j = 1$ case): When $j = 0$, according to the initialization step in Procedure 4.1,

$$P_{2,5}^{(0)} = f_{P,5}(\Lambda_2, \Gamma_2) > 0. \tag{A.58}$$

Then from Equation (A.54), we have

$$\lambda_{1,4}^{(1)} = \lambda_{1,4}^{(0)}\Big(1 - P_{2,5}^{(0)}\Big) < \lambda_{1,4}^{(0)}. \tag{A.59}$$

Following Proposition 4.3, we have

$$P_{1,5}^{(1)} = f_{P,5}\Big(\Lambda_1^{(1)}, \Gamma_1\Big) < f_{P,5}\Big(\Lambda_1^{(0)}, \Gamma_1\Big) = P_{1,5}^{(0)}. \tag{A.60}$$

In addition, we have

$$P_{1,5}^{(1)} = f_{P,5}\Big(\Lambda_1^{(1)}, \Gamma_1\Big) > 0. \tag{A.61}$$

From Equation (A.56), we have

$$\lambda_{2,4}^{(1)} = \lambda_{2,4}^{(0)}\Big(1 - P_{1,5}^{(1)}\Big) < \lambda_{2,4}^{(0)}. \tag{A.62}$$

From Equation (A.67), we have

$$P_{2,5}^{(1)} \;=\; f_{P,5}\Big(\Lambda_2^{(1)}, \Gamma_2\Big) < f_{P,5}\Big(\Lambda_2^{(0)}, \Gamma_2\Big) = P_{2,5}^{(0)}. \tag{A.63}$$

In to the next iteration, for $j = 1$, which is the base case, from Equation (A.54), we have

$$\lambda_{1,4}^{(2)} = \lambda_{1,4}^{(0)}\Big(1 - P_{2,5}^{(1)}\Big) > \lambda_{1,4}^{(0)}\Big(1 - P_{2,5}^{(0)}\Big) = \lambda_{1,4}^{(1)}. \tag{A.64}$$

Then from Proposition 4.3, we have

$$P_{1,5}^{(2)} = f_{P,5}\Big(\Lambda_1^{(2)}, \Gamma_1\Big) > f_{P,5}\Big(\Lambda_1^{(1)}, \Gamma_1\Big) = P_{1,5}^{(1)}. \tag{A.65}$$

From Equation (A.56), we have

$$\lambda_{2,4}^{(2)} = \lambda_{2,4}^{(0)}\Big(1 - P_{1,5}^{(2)}\Big) < \lambda_{2,4}^{(0)}\Big(1 - P_{1,5}^{(1)}\Big) = \lambda_{2,4}^{(1)}. \tag{A.66}$$

From Equation (A.67), we have

$$P_{2,5}^{(2)} \;=\; f_{P,5}\Big(\Lambda_2^{(2)}, \Gamma_2\Big) < f_{P,5}\Big(\Lambda_2^{(1)}, \Gamma_2\Big) = P_{2,5}^{(1)}. \tag{A.67}$$

The above proves the base case ($j = 1$).

Inductive Step: Assume when $j = k$, we have

$$\lambda_{1,4}^{(k+1)} > \lambda_{1,4}^{(k)}, \tag{A.68}$$

$$P_{1,5}^{(k+1)} > P_{1,5}^{(k)}, \tag{A.69}$$

$$\lambda_{2,4}^{(k+1)} < \lambda_{2,4}^{(k)}, \tag{A.70}$$

$$P_{2,5}^{(k+1)} < P_{2,5}^{(k)}. \tag{A.71}$$

From Equation (A.71), we have

$$\lambda_{1,4}^{(k+2)} = \lambda_{1,4}^{(0)}\Big(1 - P_{2,5}^{(k+1)}\Big) > \lambda_{1,4}^{(0)}\Big(1 - P_{2,5}^{(k)}\Big) = \lambda_{1,4}^{(k+1)}, \tag{A.72}$$

from Proposition 4.3, we have

$$P_{1,5}^{(k+2)} = f_{P,5}\Big(\Lambda_1^{(k+2)}, \Gamma_1\Big) > f_{P,5}\Big(\Lambda_1^{(k+1)}, \Gamma_1\Big) = P_{1,5}^{(k+1)}, \tag{A.73}$$

From Equation (A.69), we have

$$\lambda_{2,4}^{(k+2)} = \lambda_{2,4}^{(0)}\Big(1 - P_{1,5}^{(k+2)}\Big) < \lambda_{2,4}^{(0)}\Big(1 - P_{1,5}^{(k+1)}\Big) = \lambda_{2,4}^{(k+1)}, \tag{A.74}$$

Finally, from Proposition 4.3, we have

$$P_{2,5}^{(k+2)} = f_{P,5}\Big(\Lambda_2^{(k+2)}, \Gamma_1\Big) < f_{P,5}\Big(\Lambda_2^{(k+1)}, \Gamma_1\Big) = P_{2,5}^{(k+1)}. \tag{A.75}$$

Therefore $j = k+1$ also holds, which completes the mathematical induction. The lemma holds. ■

Proof of Proposition 4.1:

From Lemma A.5, $P_{1,5}^{(j)}$, $j = 1, 2, 3, ...$, is monotonically decreasing, also it is bounded from above and below because $0 < P_{1,5}^{(j)} < 1$, therefore $P_{1,5}^{(j)}$ is convergent. Similarly, $P_{2,5}^{(j)}$, $j = 1, 2, 3, ...$, is also bounded and monotonically increasing, so it is also convergent. Therefore we have

$$\lim_{j\to\infty} P_{k,5}^{(j)} = P_{k,5}, k = 1, 2. \tag{A.76}$$

From the balance equation in single-patient CTMC model, for patient k, we have

$$(\lambda_{k,6} + \mu_{k,4})P_{k,6} = \lambda_{k,5}P_{k,5}, k = 1, 2. \tag{A.77}$$

Since $\lambda_{k,6}$, $\mu_{k,4}$ and $\lambda_{k,5}$ are never updated, thus always stay unchanged, which make $P_{k,5}$ and $P_{k,6}$ have the same monotonic property and both are bounded. Therefore, we have

$$\lim_{j\to\infty} P_{k,6}^{(j)} = P_{k,6}, k = 1, 2. \tag{A.78}$$

This proposition is proved. ■

Proof of Proposition 4.2: This proof follows the similar logic of the proof for Proposition 4.1. ■

Appendix B: Transition rate matrix

The transition rate matrix Q of two-identical-patient CTMC model is presented below, which is the concatenation of matrices Q_1, Q_2, Q_3 and Q_4.

$Q = [Q_1 \mid Q_2 \mid Q_3 \mid Q_4]$.

$$Q_1 = \begin{pmatrix}
-2\mu_5 & 0 & 0 & 0 & 0 & 0 & 2\mu_5 \\
\lambda_6 & -\lambda_6-\mu_4-\mu_5 & 0 & 0 & 0 & 0 & \mu_4 \\
0 & \lambda_5 & -\lambda_5-\mu_3-\mu_5 & 0 & 0 & 0 & \mu_3 \\
0 & 0 & \lambda_4 & -\lambda_4-\mu_2-\mu_5 & 0 & 0 & \mu_2 \\
0 & 0 & 0 & 0 & -\mu_1-\mu_5 & 0 & \mu_1 \\
0 & 0 & 0 & \lambda_3 & 0 & -\lambda_3-\mu_5 & 0 \\
0 & 0 & 0 & 0 & \lambda_1 & \lambda_2 & -\lambda_1-\lambda_2-\mu_5 \\
0 & 0 & \lambda_6 & 0 & 0 & 0 & 0 \\
0 & \lambda_6 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda_6 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \lambda_6 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda_6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_6 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

$$Q_2 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \mu_5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-\lambda_5 - \lambda_6 - \mu_3 - \mu_4 & \lambda_5 & 0 & 0 & 0 & \mu_3 & 0 \\
0 & -\lambda_6 - \mu_4 & 0 & 0 & 0 & \mu_4 & 0 \\
\mu_4 & 0 & -\lambda_6 - \mu_2 - \mu_4 - \mu_5 & 0 & 0 & \mu_2 & 0 \\
0 & 0 & 0 & -\lambda_6 - \mu_1 - \mu_4 & 0 & \mu_1 & 0 \\
0 & 0 & \lambda_3 & 0 & -\lambda_3 - \lambda_6 - \mu_4 & 0 & 0 \\
0 & 0 & 0 & \lambda_1 & \lambda_2 & -\lambda_1 - \lambda_2 - \lambda_6 - \mu_4 & 0 \\
0 & 0 & \lambda_5 & 0 & 0 & 0 & \lambda_4 - \lambda_5 - \mu_2 - \mu_3 \\
\lambda_5 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda_5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \lambda_5 & 0 & \lambda_3 \\
0 & 0 & 0 & 0 & 0 & \lambda_5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

$$Q_3 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu_5 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \mu_5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu_4 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \mu_5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda_4 & 0 & 0 & \mu_2 & 0 & 0 & \mu_3 \\
-\mu_3 - \lambda_5 & 0 & 0 & \mu_3 & 0 & 0 & 0 \\
0 & -\lambda_5 - \mu_1 - \mu_3 & 0 & \mu_1 & 0 & 0 & 0 \\
0 & 0 & -\lambda_3 - \lambda_5 - \mu_3 & 0 & 0 & 0 & 0 \\
0 & \lambda_1 & \lambda_2 & -\lambda_1 - \lambda_2 - \lambda_5 - \mu_3 & 0 & 0 & 0 \\
0 & 0 & \lambda_4 & 0 & -\lambda_3 - \lambda_4 - \mu_2 & \lambda_3 & 0 \\
0 & 0 & 0 & 0 & 0 & -\lambda_4 - \mu_2 & \mu_2 \\
0 & 0 & 0 & \lambda_4 & \lambda_2 & 0 & -\lambda_1 - \lambda_2 - \lambda_4 - \mu_2 \\
0 & 0 & 0 & 0 & 0 & 0 & -\mu_2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \lambda_3 \\
0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

$$Q_4 = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu_5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \mu_5 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu_4 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \mu_4 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_4 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \mu_3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \mu_3 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_3 \\
0 & 0 & 0 & 0 & 0 & 0 & \mu_2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_2 \\
-\lambda_4 - \mu_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -\lambda_3 - \mu_1 & \lambda_3 & 0 & 0 & 0 & \mu_1 & 0 \\
0 & 0 & -\mu_1 & 0 & 0 & \mu_1 & 0 & 0 \\
0 & \lambda_2 & 0 & -\lambda_1 - \lambda_2 - \mu_1 & \lambda_1 & 0 & 0 & \mu_1 \\
0 & 0 & 0 & \mu_1 & -\mu_1 & 0 & 0 & 0 \\
0 & 2\lambda_3 & 0 & 0 & 0 & -2\lambda_3 & 0 & 0 \\
0 & \lambda_1 & 0 & 0 & 0 & 0 & -\lambda_1 - \lambda_3 & 0 \\
0 & 0 & 0 & \lambda_1 & 0 & 0 & \lambda_2 & -\lambda_1 - \lambda_2
\end{pmatrix}$$

Bibliography

[1] Institute for healthcare improvement. `http://www.ihi.org/IHI/programs/campaign`, May 2005.

[2] A. Ades and S. Cliffe. Markov chain monte carlo estimation of a multiparameter decision model: consistency of evidence and the accurate assessment of uncertainty. *Medical Decision Making*, 22(4):359–371, 2002.

[3] L. H. Aiken, S. P. Clarke, R. B. Cheung, D. M. Sloane, and J. H. Silber. Educational levels of hospital nurses and surgical patient mortality. *Jama*, 290(12): 1617–1623, 2003.

[4] O. Alagoz, L. M. Maillart, A. J. Schaefer, and M. S. Roberts. The optimal timing of living-donor liver transplantation. *Management Science*, 50(10):1420–1430, 2004.

[5] O. Alagoz, H. Hsu, A. J. Schaefer, and M. S. Roberts. Markov decision processes: a tool for sequential decision making under uncertainty. *Medical Decision Making*, 30(4):474–483, 2010.

[6] O. Alagoz, T. Ayer, and F. S. Erenay. Operations research models for cancer screening. *Wiley Encyclopedia of Operations Research and Management Science*, 2011.

[7] J. G. Anderson, R. Ramanujam, D. Hensel, M. M. Anderson, and C. A. Sirio. The need for organizational change in patient safety initiatives. *International Journal of Medical Informatics*, 75:809–817, 2006.

[8] T. Ayer, O. Alagoz, and N. Tout. A mathematical model to optimize breast cancer screening policy. In *Proceedings of the 31st Annual Meeting of the Society for Medical Decision Making*, pages 17–21, 2009.

[9] J. Banks, J. Carson, and B. Nelson. *DM Nicol, Discrete-Event System Simulation*. Prentice Hall, 2000.

[10] H. Beaulieu, J. A. Ferland, B. Gendron, and P. Michelon. A mathematical programming approach for scheduling physicians in the emergency room. *Health Care Management Science*, 3(3):193–200, 2000.

[11] A. Begun, A. Icks, R. Waldeyer, S. Landwehr, M. Koch, and G. Giani. Identification of a multistate continuous-time nonhomogeneous markov chain model for patients with decreased renal function. *Medical Decision Making*, 33(2):298–306, 2013.

[12] R. Bellomo, D. Goldsmith, S. Uchino, J. Buckmaster, G. Hart, H. Opdam, W. Silvester, L. Doolan, and G. Gutteridge. Prospective controlled trial of effect of medical emergency team on postoperative morbidity and mortality rates*. *Critical care medicine*, 32(4):916–921, 2004.

[13] D. M. Berwick, D. R. Calkins, C. J. McCannon, and A. D. Hackbarth. The 100000 lives campaign. *JAMA: The Journal of the American Medical Association*, 295(3): 324–327, 2006.

[14] S. Biller, J. Li, S. P. Marin, S. M. Meerkov, and L. Zhang. Bottlenecks in bernoulli serial lines with rework. *IEEE Transactions on Automation Science and Engineering*, 7(2):208–217, 2010.

[15] G. Bolch, S. Greiner, H. de Meer, and K. S. Trivedi. *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. Wiley-Interscience, 2006.

[16] A. Brahme. Development of radiation therapy optimization. *Acta Oncologica*, 39 (5):579–595, 2000.

[17] M. L. Brandeau, F. Sainfort, and W. P. Pierskalla. *Operations research and health care: a handbook of methods and applications*, volume 70. Springer, 2004.

[18] S. Brenner, Z. Zeng, Y. Liu, J. Wang, J. Li, and P. K. Howard. Modeling and analysis of the emergency department at university of kentucky chandler hospital using simulations. *Journal of Emergency Nursing*, 36(4):303, 2010.

[19] P. G. Brindley. Patient safety and acute care medicine: lessons for the future, insights from the past. *Critical Care*, 14(2):217, 2010.

[20] M. D. Buist, E. Jarmolowski, P. R. Burton, S. A. Bernard, B. P. Waxman, and J. Anderson. Recognising clinical instability in hospital patients before cardiac arrest or unplanned admission to intensive care: a pilot study in a tertiary-care hospital. *The Medical Journal of Australia*, 171(1):22, 1999.

[21] B. Cardoen, E. Demeulemeester, and J. Beliën. Operating room planning and scheduling: a literature review. *European Journal of Operational Research*, 201 (3):921–932, 2010.

[22] M. W. Carter and J. T. Blake. Using simulation in an acute-care hospital: easier said than done. In *Operations research and health care*, pages 191–215. Springer, 2004.

[23] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[24] P. S. Chan, A. Khalid, L. S. Longmore, R. A. Berg, M. Kosiborod, and J. A. Spertus. Hospital-wide code rates and mortality before and after implementation of a rapid response team. *JAMA: The Journal of the American Medical Association*, 300(21):2506–2513, 2008.

[25] P. S. Chan, R. Jain, B. K. Nallmothu, R. A. Berg, and C. Sasson. Rapid response teams: a systematic review and meta-analysis. *Archives of Internal Medicine*, 170 (1):18–26, 2010.

[26] S.-Y. Chiang, C.-T. Kuo, and S. Meerkov. c-bottlenecks in serial production lines: identification and application. *Mathematical Problems in Engineering*, 7(6):543–578, 2001.

[27] S. P. Clarke. Failure to rescue: lessons from missed opportunities in care. *Nursing inquiry*, 11(2):67–71, 2004.

[28] S. P. Clarke and L. H. Aiken. Failure to rescue: needless deaths are prime examples of the need for more nurses at the bedside. *AJN The American Journal of Nursing*, 103(1):42–47, 2003.

[29] F. C. Coelli, R. B. Ferreira, R. M. V. Almeida, and W. C. A. Pereira. Computer simulation and discrete-event models in the analysis of a mammography clinic patient flow. *Computer methods and programs in biomedicine*, 87(3):201–207, 2007.

[30] E. A. Crawford, P. J. Parikh, N. Kong, and C. V. Thakar. Analyzing discharge strategies during acute care a discrete-event simulation study. *Medical Decision Making*, page 0272989X13503500, 2013.

[31] L. Cronenwett, G. Sherwood, J. Barnsteiner, J. Disch, J. Johnson, P. Mitchell, D. T. Sullivan, and J. Warren. Quality and safety education for nurses. *Nursing outlook*, 55(3):122–131, 2007.

[32] M. J. Dacey, E. R. Mirza, V. Wilcox, M. Doherty, J. Mello, A. Boyer, J. Gates, T. Brothers, and R. Baute. The effect of a rapid response team on major clinical outcome measures in a community hospital. *Critical Care Medicine*, 35(9):2076–2082, 2007.

[33] A. M. de Bruin, A. Van Rossum, M. Visser, and G. Koole. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137, 2007.

[34] B. Denton, J. Viapiano, and A. Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health Care Management Science*, 10(1): 13–24, 2007.

[35] M. A. DeVita, R. Bellomo, K. Hillman, J. Kellum, A. Rotondi, D. Teres, A. Auerbach, W.-J. Chen, K. Duncan, and G. Kenward. Findings of the first consensus conference on medical emergency teams. *Critical Care Medicine*, 34(9):2463–2478, 2006.

[36] M. A. DeVita, G. B. Smith, S. K. Adam, I. Adams-Pizarro, M. Buist, R. Bellomo, R. Bonello, E. Cerchiari, B. Farlow, and D. Goldsmith. identifying the hospitalised patient in crisisa consensus conference on the afferent limb of rapid response systems. *Resuscitation*, 81(4):375–382, 2010.

[37] G. Dobson, H.-H. Lee, and E. Pinker. A model of icu bumping. *Operations research*, 58(6):1564–1576, 2010.

[38] A. W. Downey, J. L. Quach, M. Haase, A. Haase-Fielitz, D. Jones, and R. Bellomo. Characteristics and outcomes of patients receiving a medical emergency team review for acute change in conscious state or arrhythmias. *Critical Care Medicine*, 36(2):477–481, 2008.

[39] W. D'Souza, H. H. Zhang, D. P. Nazareth, L. Shi, and R. R. Meyer. A nested partitions framework for beam angle optimization in intensity-modulated radiation therapy. *Physics in Medicine and Biology*, 53(12):3293, 2008.

[40] S. R. Earnshaw, A. Richter, S. W. Sorensen, T. J. Hoerger, K. A. Hicks, M. Engelgau, T. Thompson, K. V. Narayan, D. F. Williamson, E. Gregg, et al. Optimal allocation of resources across four interventions for type 2 diabetes. *Medical Decision Making*, 22(suppl 1):s80–s91, 2002.

[41] D. M. Epstein, Z. Chalabi, K. Claxton, and M. Sculpher. Efficiency, equity, and

budgetary policies informing decisions using mathematical programming. *Medical Decision Making*, 27(2):128–137, 2007.

[42] S. Fomundam and J. W. Herrmann. A survey of queuing theory applications in healthcare. *Technical Report*, Institute for Systems Research, University of Maryland, College Park, MD, 2007.

[43] C. Franklin and J. Mathew. Developing strategies to prevent inhospital cardiac arrest: analyzing responses of physicians and nurses in the hours before the event. *Critical Care Medicine*, 22(2):244, 1994.

[44] P. Fung Kon Jin, M. Dijkgraaf, C. Alons, C. van Kuijk, L. Beenen, G. Koole, and J. Goslings. Improving ct scan capabilities with a new trauma workflow concept: simulation of hospital logistics using different ct scanner scenarios. *European journal of radiology*, 80(2):504–509, 2011.

[45] S. Galhotra, M. A. DeVita, R. L. Simmons, and M. A. Dew. Mature rapid response system and potentially avoidable cardiopulmonary arrests in hospital. *Quality and Safety in Health Care*, 16(4):260–265, 2007.

[46] D. Goldhill, S. White, and A. Sumner. Physiological values and procedures in the 24 h before icu admission from the ward. *Anaesthesia*, 54(6):529–534, 1999.

[47] D. Goldhill, L. Worthington, A. Mulcahy, M. Tarling, and A. Sumner. The patient-at-risk team: identifying and managing seriously ill ward patients. *ANAESTHESIA-LONDON-*, 54:853–860, 1999.

[48] L. Green. Queueing analysis in healthcare. In R. W. Hall, editor, *Patient flow: reducing delay in healthcare delivery*, pages 281–307. Springer, 2006.

[49] L. V. Green. Capacity planning and management in hospitals. In *Operations research and health care*, pages 15–41. Springer, 2004.

[50] L. V. Green. Capacity planning and management in hospitals. In M. L. Brandeau, F. Sainfort, and W. P. Pierskalla, editors, *Operations Research and Health Care*, pages 15–41. Springer, 2005.

[51] L. V. Green, S. Savin, and B. Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 54(1):11–25, 2006.

[52] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[53] N. H. S. Group. National health expenditure projections 2011-2021. `http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/Proj2011PDF.pdf`, 2011.

[54] D. Gupta and B. Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

[55] C. Haraden and R. Resar. Patient flow in hospitals: understanding and controlling it better. *Frontiers of Health Services Management*, 20(4):3–15, 2004.

[56] P. Harper and A. Shahani. Modelling for the planning and management of bed capacities in hospitals. *Journal of the Operational Research Society*, 53(1):11–18, 2002.

[57] J. E. Helm, S. AhmadBeygi, and M. P. Van Oyen. Design and analysis of hospital admission control for operational effectiveness. *Production and Operations Management*, 20(3):359–374, 2011.

[58] W. L. Herring and J. W. Herrmann. A stochastic dynamic program for the single-day surgery scheduling problem. *IIE Transactions on Healthcare Systems Engineering*, 1(4):213–225, 2011.

[59] W. L. Herring and J. W. Herrmann. The single-day surgery scheduling problem: sequential decision-making and threshold-based heuristics. *OR spectrum*, 34(2): 429–459, 2012.

[60] K. Hillman, P. Bristow, T. Chey, K. Daffurn, T. Jacques, S. Norman, G. Bishop, and G. Simmons. Antecedents to hospital deaths. *Internal Medicine Journal*, 31 (6):343–348, 2001.

[61] K. Hillman, J. Chen, M. Cretikos, R. Bellomo, D. Brown, G. Doig, S. Finfer, and A. Flabouris. Introduction of the medical emergency team (met) system: a cluster-randomised controlled trial. *Lancet*, 365(9477):2091–2097, 2005.

[62] S. H. Jacobson, S. N. Hall, and J. R. Swisher. Discrete-event simulation of health care systems. In R. W. Hall, editor, *Patient flow: reducing delay in healthcare delivery*, pages 211–252. Springer, 2006.

[63] L. Jiang and R. 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.

[64] D. Jones, R. Bellomo, and M. A. DeVita. Effectiveness of the medical emergency team: the importance of dose. *Critical Care*, 13(5):313, 2009.

[65] J. Jun, S. Jacobson, J. Swisher, et al. Application of discrete-event simulation in health care clinics: a survey. *Journal of the Operational Research Society*, 50(2): 109–123, 1999.

[66] L. C. Kaldjian, E. W. Jones, B. J. Wu, V. L. Forman-Hoffman, B. H. Levi, and G. E. Rosenthal. Reporting medical errors to improve patient safetya survey of physicians in teaching hospitals. *Archives of Internal Medicine*, 168(1):40–46, 2008.

[67] D. Kendall-Gallagher, L. H. Aiken, D. M. Sloane, and J. P. Cimiotti. Nurse specialty certification, inpatient mortality, and failure to rescue. *Journal of Nursing Scholarship*, 43(2):188–194, 2011.

[68] L. T. Kohn, J. M. Corrigan, and M. S. Donaldson. *To err is human: building a safer health system*, volume 627. National Academies Press, 2000.

[69] C.-T. Kuo, J.-T. Lim, and S. Meerkov. Bottlenecks in serial production lines: A system-theoretic approach. *Mathematical Problems in Engineering*, 2(3):233–276, 1996.

[70] C. P. Landrigan, J. M. Rothschild, J. W. Cronin, R. Kaushal, E. Burdick, J. T. Katz, C. M. Lilly, P. H. Stone, S. W. Lockley, and D. W. Bates. Effect of reducing interns' work hours on serious medical errors in intensive care units. *New England Journal of Medicine*, 351(18):1838–1848, 2004.

[71] L. L. Leape and D. M. Berwick. Five years after to err is human. *JAMA: The Journal of the American Medical Association*, 293(19):2384–2390, 2005.

[72] E. K. Lee, T. Fox, and I. Crocker. Integer programming applied to intensity-modulated radiation therapy treatment planning. *Annals of Operations Research*, 119(1-4):165–181, 2003.

[73] J. Li. Throughput analysis in automotive paint shops: a case study. *IEEE Transactions on Automation Science and Engineering*, 1(1):90–98, 2004.

[74] J. Li and S. Meerkov. *Production systems engineering*. Springer, 2009.

[75] E. Litvak and P. J. Pronovost. Rethinking rapid response teams. *JAMA: The Journal of the American Medical Association*, 304(12):1375–1376, 2010.

[76] A. Mandelbaum, P. Momcilovic, and Y. Tseytlin. On fair routing from emergency departments to hospital wards: Qed queues with heterogeneous servers. *Management Science*, 58(7):1273–1291, 2012.

[77] D. Massey, L. M. Aitken, and W. Chaboyer. Literature review: do rapid response systems reduce the incidence of major adverse events in the deteriorating ward patient? *Journal of clinical nursing*, 19(23-24):3260–3273, 2010.

[78] F. McArthur-Rouse. Critical care outreach services and early warning scoring systems: a review of the literature. *Journal of Advanced Nursing*, 36(5):696–704, 2001.

[79] J. McGaughey, F. Alderdice, R. Fowler, A. Kapila, A. Mayhew, and M. Moutray. Outreach and early warning systems (ews) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database of Systematic Reviews*, 3, 2007.

[80] H. McGloin, S. K. Adam, and M. Singer. Unexpected deaths and referrals to intensive care of patients on general wards. are some cases potentially avoidable? *Journal of the Royal College of Physicians of London*, 33(3):255–259, 1998.

[81] P. McQuillan, S. Pilkington, A. Allan, B. Taylor, A. Short, G. Morgan, M. Nielsen, D. Barrett, and G. Smith. Confidential inquiry into quality of care before admission to intensive care. *Bmj*, 316(7148):1853–1858, 1998.

[82] R. R. Meyer, H. H. Zhang, L. Goadrich, D. P. Nazareth, L. Shi, and W. D. D-Souza. A multiplan treatment-planning framework: a paradigm shift for intensity-modulated radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, 68(4):1178–1189, 2007.

[83] K. Muthuraman and M. Lawley. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *Iie Transactions*, 40(9):820–837, 2008.

[84] J. Needleman, P. Buerhaus, S. Mattke, M. Stewart, and K. Zelevinsky. Nurse-staffing levels and the quality of care in hospitals. *New England Journal of Medicine*, 346(22):1715–1722, 2002.

[85] J. P. Oddoye, D. F. Jones, M. Tamiz, and P. Schmidt. Combining simulation and goal programming for healthcare planning in a medical assessment unit. *European Journal of Operational Research*, 193(1):250–261, 2009.

[86] M. A. Peberdy, M. Cretikos, B. S. Abella, M. DeVita, D. Goldhill, W. Kloeck, S. L. Kronick, L. J. Morrison, V. M. Nadkarni, G. Nichol, et al. Recommended guidelines for monitoring, reporting, and conducting research on medical emergency team, outreach, and rapid response systems: An utstein-style scientific statement a scientific statement from the international liaison committee on resuscitation (american heart association, australian resuscitation council, european resuscitation council, heart and stroke foundation of canada, interamerican heart foundation, resuscitation council of southern africa, and the new zealand resuscitation council); the american heart association emergency cardiovascular care committee; the council on cardiopulmonary, perioperative, and critical care; and the interdisciplinary working group on quality of care and outcomes research. *Circulation*, 116(21): 2481–2500, 2007.

[87] E. Pinker and T. Tezcan. Determining the optimal configuration of hospital inpatient rooms in the presence of isolation patients. *Operations Research*, 61(6): 1259–1276, 2013.

[88] G. Priestley, W. Watson, A. Rashidian, C. Mozley, D. Russell, J. Wilson, J. Cope, D. Hart, D. Kay, and K. Cowley. Introducing critical care outreach: a ward-randomised trial of phased introduction in a general hospital. *Intensive Care Medicine*, 30(7):1398–1404, 2004.

[89] S. R. Ranji, A. D. Auerbach, C. J. Hurd, K. O'Rourke, and K. G. Shojania. Effects of rapid response systems on clinical outcomes: systematic review and meta-analysis. *Journal of Hospital Medicine*, 2(6):422–432, 2007.

[90] M. Reynolds, C. Vasilakis, M. McLeod, N. Barber, A. Mounsey, S. Newton, A. Jacklin, and B. D. Franklin. Using discrete event simulation to design a more efficient hospital pharmacy for outpatients. *Health care management science*, 14 (3):223–236, 2011.

[91] R. T. Ridley. The relationship between nurse education level and patient safety: an integrative review. *The Journal of nursing education*, 47(4):149–156, 2008.

[92] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, A. Kumar, and J. G. Li. A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planning. *Physics in Medicine and Biology*, 48(21): 3521, 2003.

[93] A. Schmid, L. Hoffman, M. B. Happ, G. A. Wolf, and M. DeVita. Failure to rescue: a literature review. *Journal of Nursing Administration*, 37(4):188–198, 2007.

[94] F. Sebat, A. A. Musthafa, D. Johnson, A. A. Kramer, D. Shoffner, M. Eliason, K. Henry, and B. Spurlock. Effect of a rapid response system for patients in shock on time to treatment and mortality during 5 years*. *Critical care medicine*, 35 (11):2568–2575, 2007.

[95] X. Shao, J. Li, and D. A. Wiegmann. Bottleneck analysis to reduce flow disruptions on surgery: theory and application. *Technical Report*, Department of Industrial and Systems Engineering, University of Wisconsin, Madison, WI, 2013.

[96] P. J. Sharek, L. M. Parast, K. Leong, J. Coombs, K. Earnest, J. Sullivan, L. R. Frankel, and S. J. Roth. Effect of a rapid response team on hospital-wide mortality and code rates outside the icu in a childrens hospital. *JAMA: The Journal of the American Medical Association*, 298(19):2267–2274, 2007.

[97] K. G. Shojania, B. W. Duncan, K. M. McDonald, R. M. Wachter, and A. J. Markowitz. *Making health care safer: a critical analysis of patient safety practices*. Agency for Healthcare Research and Quality Rockville, MD, 2001.

[98] J. H. Silber, S. V. Williams, H. Krakauer, and J. S. Schwartz. Hospital and patient characteristics associated with death after surgery: a study of adverse occurrence and failure to rescue. *Medical care*, pages 615–629, 1992.

[99] J. H. Silber, P. R. Rosenbaum, and R. N. Ross. Comparing the contributions of groups of predictors: which outcomes vary with hospital rather than patient characteristics? *Journal of the American Statistical Association*, 90(429):7–18, 1995.

[100] J. H. Silber, P. S. Romano, A. K. Rosen, Y. Wang, O. Even-Shoshan, and K. G. Volpp. Failure-to-rescue: comparing definitions to measure quality of care. *Medical care*, 45(10):918–925, 2007.

[101] M. D. Silverstein, E. V. Loftus, W. J. Sandborn, W. J. Tremaine, B. G. Feagan, P. J. Nietert, W. S. Harmsen, and A. R. Zinsmeister. Clinical course and costs of care for crohn's disease: Markov model analysis of a population-based cohort. *Gastroenterology*, 117(1):49–57, 1999.

[102] A. F. Smith and J. Wood. Can some in-hospital cardio-respiratory arrests be prevented? a prospective survey. *Resuscitation*, 37(3):133–137, 1998.

[103] C. W. Spry and M. A. Lawley. Evaluating hospital pharmacy staffing and work scheduling using simulation. In *Proceedings of the 37th conference on Winter simulation*, pages 2256–2263. Winter Simulation Conference, 2005.

[104] A. H. Taenzer, J. B. Pyke, and S. P. McGrath. A review of current and emerging approaches to address failure-to-rescue. *Anesthesiology*, 115(2):421–431, 2011.

[105] S. Taylor, T. Eldabi, G. Riley, R. Paul, and M. Pidd. Simulation modelling is 50! do we need a reality check? *Journal of the Operational Research Society*, 60 (Special):S69–S82, 2009.

[106] R. M. Trinkle and A. Flabouris. Documenting rapid response system afferent limb failure and associated patient outcomes. *Resuscitation*, 82(7):810–814, 2011.

[107] R. M. Wachter. The end of the beginning: patient safety five years after to err is human. *Health Affairs*, 23(11):534–545, 2004.

[108] J. Wang, J. Li, K. Tussey, and K. Ross. Reducing length of stay in emergency department: A simulation study at a community hospital. *IEEE Transactions on Systems, Man and Cybernetics, part A: Systems and Humans*, 42(6):1314–1322, 2012.

[109] J. Wang, S. Quan, J. Li, and A. 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.

[110] J. Wang, J. Li, and P. K. Howard. A system model of work flow in the patient room of hospital emergency department. *Health Care Management Science*, DOI: 10.1007/s10729-013-9235-1, 2013.

[111] J. Wang, X. Zhong, J. Li, and P. K. Howard. Modeling and analysis of care delivery services within patient rooms: a system-theoretic approach. *IEEE Transactions on Automation Science and Engineering*, DOI: 10.1109/TASE.2013.2242326, 2013.

[112] E. N. Weiss, M. A. Cohen, and J. C. Hershey. An iterative estimation and validation procedure for specification of semi-markov models with application to hospital patient flow. *Operations Research*, 30(6):1082–1104, 1982.

[113] N. J. Welton and A. Ades. Estimation of markov chain transition probabilities and rates from fully and partially observed data: uncertainty propagation, evidence synthesis, and model calibration. *Medical Decision Making*, 25(6):633–645, 2005.

[114] J. L. Wiler, R. T. Griffey, and T. Olsen. Review of modeling approaches for emergency department patient flow and crowding research. *Academic Emergency Medicine*, 18(12):1371–1379, 2011.

[115] B. D. Winters, J. C. Pham, E. A. Hunt, E. Guallar, S. Berenholtz, and P. J. Pronovost. Rapid response systems: a systematic review. *Critical Care Medicine*, 35(5):1238–1243, 2007.

[116] B. D. Winters, P. J. Pronovost, M. Miller, and E. A. Hunt. Measuring and improving safety. In M. A. DeVita, editor, *Textbook of Rapid Response Systems*, pages 19–35. Springer, 2011.

[117] N. Yankovic and L. V. Green. Identifying good nursing levels: A queuing approach. *Operations research*, 59(4):942–955, 2011.

[118] Z. Zeng, X. Xie, X. Zhong, J. Li, B. A. Liegel, and S. Sanford-Ring. Simulation modeling of hospital discharge process. In *Proceedings of the 2013 Industrial and Systems Engineering Research Conference, San Juan, Puerto Rico, May 2013*.

[119] Z. Zeng, X. Ma, Y. Hu, J. Li, and D. 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.

[120] H. H. Zhang, L. Shi, R. Meyer, D. Nazareth, and W. D'Souza. Solving beam-angle selection and dose optimization simultaneously via high-throughput computing. *INFORMS Journal on Computing*, 21(3):427–444, 2009.

[121] Y. Zhang, M. L. Puterman, M. Nelson, and D. Atkins. A simulation optimization approach to long-term care capacity planning. *Operations Research*, 60(2):249–261, 2012.

[122] X. Zhong, M. Williams, J. Li, and S. Kraft. Primary care redesign: A simulation study at a pediatric clinic. In *Proceedings of IEEE conference on Automation Science and Engineering, Taiwan, Aug 2014*.