March 2022

Statistical Monitoring the Quality of Healthcare Services

Yanqing Kuang
University of South Florida

Statistical Monitoring the Quality of Healthcare Services

by

Yanqing Kuang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Co-Major Professor: Devashish Das, Ph.D.
Co-Major Professor: Mingyang Li, Ph.D.
Tapas K. Das, Ph.D.
Yasin Yilmaz, Ph.D.
Lu Lu, Ph.D.

Date of Approval:
March 28, 2022

Keywords: Statistical Monitoring Methods, Healthcare Service Quality, Stochastic Process Models, Emergency Department, Alcohol Use Disorder

Dedication

I dedicate this dissertation to my husband, my daughter, my parents and to all who helped me along the way.

Acknowledgments

First, I would like to express my deepest gratitude to my co-major advisor, Dr. Devashish Das, for his dedicated support and professional guidance throughout my doctoral studies. It is his guidance, encouragement, sincerity that makes me go through those hard times. This dissertation would not have become possible without his valuable supervision and advice.

Furthermore, I would like to show my sincere gratitude to my co-major advisor, Dr. Mingyang Li, for his valuable support and advice in improving my dissertation. He continuously provided encouragement and was always willing and enthusiastic to assist in any way he could throughout this dissertation.

Also, I would like to acknowledge my sincere thanks to my advisory committee members, Dr. Tapas K. Das, Dr. Yasin Yilmaz and Dr. Lu Lu for their time and insightful suggestions for my research. I would like to thank all my colleagues and friends at the University of South Florida who have helped me so much in my studies and daily life.

Last but not the least, I am extremely grateful to my husband for his love, understanding, and continuing supports in my doctoral studies. I am very much thankful to my parents for their love, caring, and sacrifices for educating me for my future.

Table of Contents

List of Tables iii

List of Figures iv

Abstract vii

Chapter 1: Introduction 1
- 1.1 Background and Motivation 1
- 1.2 Literature Review 3
  - 1.2.1 Statistical Process Control Methods for Service Processes . . 3
  - 1.2.2 Statistical Monitoring the Quality of Acute Care Service . . 7
  - 1.2.3 Statistical Monitoring the Quality of Chronic Care Service . 9
- 1.3 Overview and Organization of the Dissertation 12

Chapter 2: Monitoring Timeliness of Healthcare Delivery in Emergency Department Using Counting Processes 16
- 2.1 Overview 16
- 2.2 Introduction 17
- 2.3 Quadratic Contrast Estimation 20
- 2.4 Quadratic Contrast Tests 29
- 2.5 Simulation Study 32
  - 2.5.1 Detecting Changes in Departure Rate of Single-server Queues 34
    - 2.5.1.1 $M_t/M_t/1$ Queue 34
    - 2.5.1.2 $M_t/G_t/1$ Queue 35
    - 2.5.1.3 $G_t/G_t/1$ Queue 38
  - 2.5.2 Detecting Changes in Departure Rate of Multi-server Queues 39
- 2.6 Monitoring the Waiting Patients Waiting Volume in an Hospital ED 43
- 2.7 Conclusions 46

Chapter 3: Statistical Monitoring of the Quality of Service in a Network of Queues with Application in Emergency Department 47
- 3.1 Overview 47
- 3.2 Introduction 47
- 3.3 Queuing Network Model 51
- 3.4 Statistical Monitoring Scheme for Queueing Network 53
- 3.5 Proposed CUSUM Charts 54
  - 3.5.1 The SCUSUM Chart 55

3.5.2 The G-CUSUM and P-CUSUM Charts . . . 56
3.6 Numerical Study . . . 58
3.6.1 ARL Comparisons for Detecting the Change of All Service Nodes . . . 60
3.6.2 ARL Comparisons for Detecting the Change of Single Node 61
3.6.3 Identity the Exact Out-of-control Node Using Penalized CUSUM Chart . . . 63
3.7 Case Study: Monitoring the Flow of Patients in an ED . . . 65
3.8 Conclusion . . . 71

Chapter 4: Statistical Monitoring the Cascade of Care for Patients with Alcohol Use Disorder . . . 73
4.1 Overview . . . 73
4.2 Introduction . . . 74
4.3 Continuous-time COC (CTCOC) Model . . . 76
4.4 Statistical Monitoring Scheme . . . 78
4.5 Simulation Study . . . 80
4.6 Real Case Study . . . 83
4.6.1 Data Description . . . 83
4.6.2 Detect the Patients with Undesirable Outcomes . . . 84
4.7 Identify Key Factors Affecting the Patient Outcome . . . 85
4.7.1 Machine Learning Methods . . . 86
4.7.2 Model Performance Comparison . . . 93
4.7.3 Feature Importance . . . 97
4.8 Conclusion . . . 99

Chapter 5: Conclusion and Future Work . . . 100

References . . . 103

Appendix A: Copyright Permission . . . 119

Appendix B: Supplemental Materials . . . 120
B.1 Appendix for Chapter 2 . . . 120
B.1.1 Penalized GLR for Poisson Process . . . 120

List of Tables

Table 2.1 Confusion matrix for SQCT and ALOS chart 44

Table 2.2 Confusion matrix for GQCT and ALOS chart 45

Table 3.1 The accuracy of P-CUSUM for identifying node 2 as out-of-control if only node 2 has decreased 64

Table 3.2 CV errors for different models 68

Table 3.3 Confusion matrix for SCUSUM and MEWMA charts 69

Table 3.4 Confusion matrix for P-CUSUM and MEWMA charts 69

Table 3.5 Confusion matrix for SCUSUM and MCUSUM charts 69

Table 3.6 Confusion matrix for P-CUSUM and MCUSUM charts 69

Table 3.7 The average queue length comparisons on October 20, 2016 71

Table 4.1 Details of model parameter settings in performance comparison analysis 95

Table 4.2 Model performance comparison 96

List of Figures

Figure 1.1 Diagram of the patient treatment trajectory (Source: Duan et al. (2019)) 2

Figure 1.2 Organization of dissertation 15

Figure 2.1 Illustration of stochastic processes: (a) Time-homogeneous stochastic process (b) Time-inhomogeneous stochastic process 18

Figure 2.2 Illustration of the proposed idea for monitoring the departures in a queue: (a) The timestamps of the departures (b) The departure intensity is monitored using a counting process that denotes the number of departures in [0,t], which is denoted as $N(t)$. 19

Figure 2.3 Detecting increase in ρ of a $M_t/M_t/1$ queue. 35

Figure 2.4 Detecting change in ω of a $M_t/M_t/1$ queue. 36

Figure 2.5 Detecting increase in ρ of a $M_t/G_t/1$ queue. 37

Figure 2.6 Detecting change in ω of a $M_t/G_t/1$ queue. 37

Figure 2.7 Detecting increase in ρ of a $G_t/G_t/1$ queue. 38

Figure 2.8 Detecting change in ω of a $G_t/G_t/1$ queue. 39

Figure 2.9 Detecting increase in ρ of a $M_t/M_t/5$ queue. 40

Figure 2.10 Detecting change in ω of a $M_t/M_t/5$ queue. 40

Figure 2.11 Detecting increase in ρ of a $M_t/G_t/5$ queue. 41

Figure 2.12 Detecting change in ω of a $M_t/G_t/5$ queue. 41

Figure 2.13 Detecting increase in ρ of a $G_t/G_t/5$ queue. 42

Figure 2.14 Detecting change in ω of a $G_t/G_t/5$ queue. 42

Figure 2.15 The ED patient flow from door to bed 43

Figure 2.16 In-control bed assignment rate 44

Figure 2.17 Intensity comparison on August 26, which was signaled out-of-control by both SQCT and GQCT but not the ALOS chart 45

Figure 3.1 Structure of QN 59

Figure 3.2 ARL comparisons in detecting the decrease of the service rates of all nodes 61

Figure 3.3 ARL comparisons in detecting the decrease of μ_1 62

Figure 3.4 ARL comparisons in detecting the decrease of μ_{10} 63

Figure 3.5 ARL comparisons in detecting the decrease of μ_5 64

Figure 3.6 Parallel queueing network 65

Figure 3.7 Patient visit flow of the emergency department (ED) of a large academic medical center 66

Figure 3.8 Histogram of patient occupancy of the east node in ED 67

Figure 3.9 East node departure intensity comparison on Oct 20, 2016, which was signaled out-of-control by both SCUSUM and P-CUSUM but not the MEWMA and MCUSUM 70

Figure 4.1 CTCOC model for COC event log 77

Figure 4.2 Rate of occurrence of E_q after E_p 78

Figure 4.3 Measure of deviation from ideal COC 80

Figure 4.4 The type II error rates for detecting increase in θ 82

Figure 4.5 The type II error rates for detecting decrease in θ 82

Figure 4.6 The boxplot of the days spent for the follow-up visit after initiating the treatment for all in-control data 84

Figure 4.7 The rate of occurrence for the follow-up visit after treatment is initiated for all in-control data 84

Figure 4.8 The visit trajectory for a patient who is identified as out-of-control by our proposed method but not the traditional method 85

Figure 4.9 Typical kernel functions used in Support Vector Machines 88

Figure 4.10 The LDA computation steps 89

Figure 4.11 Patient factors that may affect the patient outcome 93

Figure 4.12 Correlation matrix between features 96

Figure 4.13 Patient's feature importances in affecting treatment outcome 97

Figure 4.14 Boxplot of the out-of-pocket medical expenses 98

Figure B.1 Detecting decrease in λ for a simple Possion process using log-likelihood ratio test 121

Abstract

In today's healthcare industry, quality of care is a growing focus in the delivery of healthcare. To improve the quality of care in healthcare delivery, many studies focus on the long-term operational decision making to meet the expectations of healthcare providers and users, such as medical resource allocation, bed planning, staff scheduling, etc. These problems are typically parts of long-term operational decision making, however, time is essential in healthcare system. To ensure the adherence to a high quality of care and detect deterioration in real time, the quality of service should be measured over days or hours instead of just months or years. Therefore, it is critical to develop effective statistical monitoring methods for detecting the deterioration in the quality of healthcare services. In this dissertation, a series of statistical monitoring methods based on stochastic process models are developed for improving the service quality in healthcare including acute and chronic care services. First, a novel statistical monitoring method based on quadratic contrast estimation technique is proposed for detecting changes in the departure intensity function in emergency department. The proposed method is based on an approximate likelihood function that alleviates the issue of needing to numerically maximize a complex likelihood function for estimating the in-control parameters and obtaining test statistics. Second, likelihood-ratio based cumulative sum (CUSUM) control charts are proposed for monitoring the service rate of queueing network with time-inhomogenerous state dependent queues. The proposed approaches could overcome the limitation of the normality assumption of traditional multivariate control charts and do not need to know the potential change in service rate of the queueing nodes in a queueing network, and thus have important practical applications. Third, a continuous-time stochastic process model is proposed to monitor and measure the treatment process for patients with alcohol use disorder (AUD) based on the Cascade of care (COC) framework. The

proposed work learns the ideal patterns in the initiation and duration of AUD treatment, from which benchmarks for COC can be developed and factors that are correlated to undesirable patient outcomes identified. Simulation studies and real case studies are considered to illustrate the proposed statistical monitoring methods and demonstrate their superior performance over traditional methods.

Chapter 1: Introduction

1.1 Background and Motivation

As today's US healthcare industry is shifting from a fee-for-service system to a value-based care system where healthcare providers are rewarded based on quality rather than quantity, quality assurance and improvement is a growing focus in the delivery of healthcare services. There are many concepts and definitions regarding the quality in healthcare. The most widely accepted definition to healthcare quality is "the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge" [1], which is first proposed by the Institute of Medicine (IOM) in 1990. A comprehensive review about the features and dimensions in the context of healthcare service quality is summarized in [2]. This paper highlighted the key characteristics of quality in healthcare delivery including safety, timeliness, effectiveness, efficiency, equity, and patient centeredness. To improve the quality of care in healthcare delivery, many studies focus on the long-term operational decision making to meet the expectations of healthcare providers and users, such as medical resource allocation/planning, bed planning, staff scheduling. These problems are typically part of long-term operational decision making, however, time is essential in healthcare system. To ensure the adherence to a high quality of care and detect deterioration in real time, the quality of service should be measured over days or hours instead of just months or years. Therefore, it is critical to develop effective statistical monitoring methods that quickly detect deterioration in the performance of the quality indicators in healthcare from which the data is collected.

Longitudinal time-to-event data is the most common type of data encountered in healthcare system. In longitudinal healthcare studies, patients are observed over time and a se-

quence of clinical events are collected on several occasions. Such examples include death, transition or recurrence of a disease, arrival or departure from a hospital, treatment trajectories for patients, etc. Figure 1.1 is an example of the treatment trajectories for patients that obtained from Electronic Health Record (EHR) [3]. In the figure, different colors and shapes represent the occur times of different clinical events, such as the diagnosis of a disease, first follow-up with the physician, first visit to physical therapy. Additionally, longitudinal time-to-event data are commonly collected as performance indicator metrics. A few representative examples include - (1) study of manufacturing processes where the quality of each finished product is assigned a label, such as "acceptable", "repair", or "discard"; (2) the number of customers waiting in a call center at any given time point of operation; (3) utilization metrics of hospital emergency departments (ED), which can be labeled as "not busy", "moderately busy", and "highly busy" during different times of the day. These performance metrics provide important insight to engineers, operations managers and healthcare providers regarding the quality of service processes. However, existing papers focus on advancing the performance modeling rather than performance monitoring in health care using the quality data that represented by these longitudinal categorical and count time-to-event data.

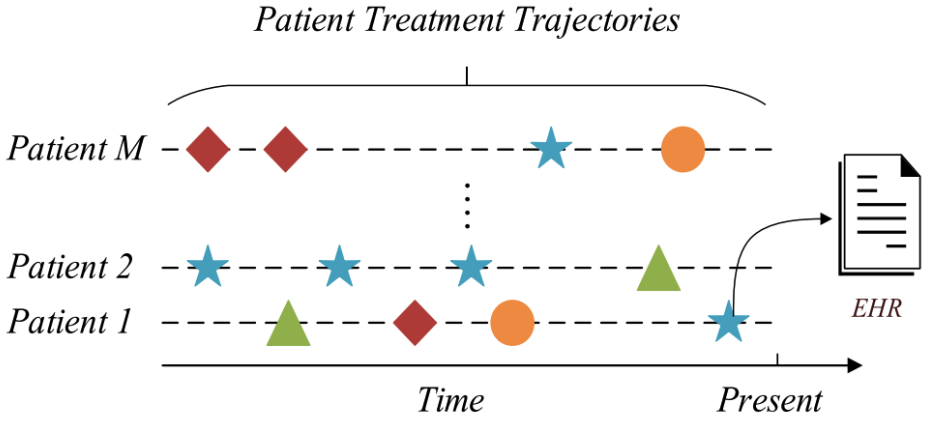


Figure 1.1: Diagram of the patient treatment trajectory (Source: Duan et al. (2019))

With the need to advance the performance monitoring in health systems engineering using time-to-event data, the main challenges are posed by stochastic process modeling and

how to effectively and accurately monitor the system performance in real time. Longitudinal time-to-event data is naturally counting process since it is generated by observing repeated measurements on a number of events that can be occurred over time. However, stochastic processes like counting processes are rarely integrated into statistical monitoring methods. Moreover, traditional statistical monitoring methods often have underlying assumption to the data set and do not fit real world data, which make the performance of traditional statistical monitoring methods inferior. Facing these challenges, this dissertation proposes novel statistical monitoring methods to effectively detect deterioration in healthcare quality based on stochastic process models.

1.2 Literature Review

In this section, comprehensive SPC methods for quality monitoring and improvement in service systems are first summarized in subsection 1.2.1. Although SPC methods have been widely used in service industries, their application in healthcare environments has not been well explored. Then, literature reviews of existing statistical monitoring studies for two major types of care services in healthcare systems engineering, namely acute care service and chronic care service, are presented in subsections 1.2.2 and 1.2.3.

1.2.1 Statistical Process Control Methods for Service Processes

Statistical Process Control (SPC) methods have been used to monitor service processes with the objective to maintain and improve the quality of service in many service systems including healthcare systems, transportation systems, and computer networks. Based on the quality related data often encountered in service system, we review SPC methods for service processes with continuous data and categorical/count data.

Many conventional univariate and multivariate control charts are used to deal with service processes with continuous data. Sulek et al. [4] employed X chart to detect the service process instability with the objective of improving the service quality and performance for a food

retailing company. Apte and Reynolds [5] presented r-bar chart and x-bar charts to monitor the window hang time to ensure the stability of a service system. Shafqat [6] designed an X-bar control chart with inverse rayleigh distribution based on repetitive sampling scheme. Mehring [7] developed a new statistical process control method to monitor the timeliness in service for a credit company to improve their customer satisfaction. The authors in [8] applied XmR chart to real-time monitor the care delivery process for outpatients in behavioral healthcare organizations. Costa and Rahim [9] proposed a synthetic control chart with non-central chi-square statistic to monitor the process mean and variance, which outperforms the X bar and R charts. Ajadi et al. [10] developed a univariate control chart called progressive mean EWMA control chart that can effectively detect small and moderate shifts in process mean. In addition, Woodall [11] presented a detailed review of the control charts methods that have been used in healthcare and public healthcare engineering.

On the other hand, when multiple dimensions of the service quality are identified, univariate control charts focusing on single continuous variable are not sufficient for service processes that are represented by several continuous variables. For example, a hospital has multiple wards to serve different types of patients and a patient may flow from one ward to another ward. The service quality measurement then can be characterized by the service rate of patients that been treated at different wards and the length of stay at different wards. It is also important to consider the relationships between different variables. For this example, the length of stay for each patient would increase when the number of patients increases, which delays the treatment for each patient. Therefore, multivariate control charts are needed to develop to monitor the multiple variables and their relationship to ensure the service quality. Jensen and Markland [12] proposed a quality perception control chart to manage the SERVQUAL quality data which are characterized by five dimensions, called "tangibles, reliability, responsiveness, assurance, empathy." Aparisi et al. [13] developed Hotelling's T^2 chart based on multiple variables supplemented with runs rules to improve the performance in detecting small or moderate shifts in process. Later, Ghute and Shirke [14] developed

a synthetic T^2 chart to monitor the mean of a multivariate normally distributed process, which integrated the Hotelling's T^2 chart and conforming run length chart. They demonstrated that their method is better than conventional Hotelling's T^2 chart and T^2 chart with supplementary runs rules. However, these Hotelling's T^2 charts all assume the successive observation data are independent, which cannot fit the autocorrelation data in practice. Thus Dargopatil and Ghute [15] designed a synthetic T^2 chart to monitor the bivariate process when variables and observations are correlated. This method combined the Hotelling's T^2 chart and the conforming run length chart, and various sampling strategies are introduced to improve the performance of the synthetic T^2 chart. Recently, Hadian and Rahimifard [16] proposed a multivariate statistical control chart for monitoring the project duration and cost based on the earned value management indices. Khae et al. [17] developed a novel synthetic multivariate control chart technique to monitor the coefficient of variation. Samanta and Mondal [18] evaluated multiple multivariate control charts for monitoring the online process in industrial engineering. He et al. [19] proposed real-time contrasts (RTC) control charts to monitor the changes in multivariate processes based on support vector machines technique. Also, multivariate cumulative sum (MCUSUM) and multivariate exponentially weighted moving average (MEWMA) charts were developed for monitoring service processes with multiple dimensions. For example, Mehmood et al. [20] established MCUSUM control chart based on bivariate ranked set schemes for process capability monitoring. Xie et al. [21] proposed a MCUSUM control chart focusing on detecting the shift for Gumbel's bivariate exponential data. Majika et al. [22] designed multivariate triple EWMA control chart for monitoring the system parameters, which is shown to be more sensitive than multivariate simple and double EWMA charts. Ajadi et al. [23] proposed a novel MEWMA for monitoring the process dispersion, which is shown to be robust to data that violates the normality assumption.

Other than the continuous data, categorical or count data are the most common type of data encountered in service processes. Many statistical process control methods are de-

veloped for monitoring the quality of service processes with categorical data. For example, Bourke [24] proposed a conforming run length (CRL) chart to monitor the change in fraction defective for sampling inspection. Wu et al. [25] integrated the CRL chart with np chart to detect the nonconforming fraction increase. Later on, Gadre and Rattihalli [26] developed a unit and group-Runs Chart which outperforms the CRL chart, the np chart and the synthetic chart in detecting fraction nonconforming increase. In addition, Li et al. [27] suggested a simple categorical control chart based on ordinal information for monitoring the attribute level count data. Jin and Loosveldt [28] designed a nonparametric multivariate statistical process control tool based on principal component analysis mix method to monitor the categorical variables. However, these control charts for categorical data were all designed for monitoring single process and they did not consider multiple stages of a service process. While considering multiple stages in a service process with categorical data, one stage may affect the performance of next stage, thus the statistical monitoring methods focusing on individual stage are not appropriate to identify the abnormality for a multiple-stage process. Sulek et al. [29] proposed a regression-based control chart, called the cause selecting control chart, to monitor a multistage service process with the cascade property. Skinner et al. [30] developed a control chart based on generalized linear model, one type of likelihood ratio statistic, to monitor the multiple discrete count data. Sogandi et al. [31] suggested a risk-adjusted control chart to control the healthcare service processes with multiple stages and categorical covariates. They showed that the likelihood ratio test is a promising method in effectively detecting the deterioration in multistage service processes. Thus, as an extension, I propose novel statistical monitoring methods by integrating the likelihood ratio test to monitor the healthcare system with multiple stages. Although SPC methods have been widely used in service industries, their application in healthcare environments has not been well explored. Then literature reviews of existing statistical monitoring studies for two major types of care services in healthcare systems engineering, namely acute care service and chronic care service, are presented in the next two subsections.

1.2.2 Statistical Monitoring the Quality of Acute Care Service

As populations continue to grow and age, acute care service in healthcare is increasingly important to respond to sudden, often unexpected, urgent or emergent injury and illness that can lead to death or disability without prompt intervention [32]. Thus monitoring the acute care service is imperative to maintain the quality and safety in healthcare delivery, and prevent serious consequences caused by potential anomalies during care process. Among all the populations, elderly patients and pediatric patients are the frailest patients who need extra attention. Brand [33] developed quality indicators to monitor the outcomes of aged people under acute care setting. Baldewijns et al. [34] proposed three techniques, tabular CUSUM, standardized CUSUM and EWMA control charts, for automatically detecting health changes in older adults based on transfer times. Khandoker et al. [35] developed a support vector machine model based on wavelet analysis to identify the older adults with a high risk of falls and injuries. Another time and energy-saving fall detection method for elderly people based on a cumulative sum control chart is introduced by Thammasat and Chaicharn [36]. Ranhoff et al. [37] applied MNA-SF method to identify the malnutrition in elderly acute medical patients. In pediatric acute healthcare settings, multiple papers investigated the effectiveness of using control charts to monitor care process. Desa et al. [38] proposed a residual control chart to monitor the performance of pediatrics hospital admission, which demonstrated the effectiveness of using pre-whitening technique for auto-correlated data. Hsian et al. [39] applied the percent coefficient of variation to detect the nonadherence and acute rejection in pediatric kidney transplant patients. Arienzo et al. [40] employed statistical analysis methods such as student t tests, Mann-Whitney U tests and chi-square tests to identify the acute kidney injury in children admitted to the pediatric intensive care unit and evaluate its influence on the outcomes in children. Moreover, Moss et al. [41] proposed an advanced time series analysis-based method to monitor the cardiorespiratory dynamics data for improving the performance for detecting the anomalies in acute care patients.

Other than the patient-level monitoring studies, system-level monitoring of acute care service is another important part needed to be investigated in this dissertation. The main acute care sectors in healthcare are intensive care unit(ICU) and emergency department (ED), thus systematic monitoring of service quality in ICU and EDs is a critical issue worldwide. Cook proposed [42] a risk adjusted statistical monitoring control chart based on machine learning methods, such as artificial neural networks (ANNs) and support vector machines (SVMs), to monitor the in-hospital mortality rate for ICU patients. Similarly, Cook et al. [43] designed a modified risk adjusted Shewhart p chart and cumulative sum process control chart to monitor the quality process and patient outcomes in ICU. As an extension, Koetsier [44] compared the performance of different risk-adjusted(RA) control charts including RA P-chart, RA Additive P-chart, RA Multiplicative P-chart, RA CUSUM, RA Resetting Sequential Probability Ratio Test, and RA EWMA control chart. They found the RA EWMA control chart performs best in detecting the change of the mortality rate of ICU patients. Rodrigues et al. [45] applied CUSUM chart to identify the risk of the prevalence of multidrug-resistant bacteria in the ICU. Cocanour et al. [46] employed a control chart for reducing the ventilator-associated pneumonia in a shock trauma ICU. Medlock et al. [47] developed statistical process control charts to monitor the timeliness of discharge letter based on the mean time elapsed between discharge and the finalized intensive care unit discharge letter, which is shown to be a multifaceted intervention that can be highly effective for improving discharge communication from the ICU.

Furthermore, statistical process control (SPC) methods have been studied in the context of monitoring the quality of service in the ED [48]. Kadri et al. [49] presented a time-series analysis model-based SPC control chart, called stationary auto-regression moving average based EWMA chart, to monitor the abnormal situation caused by the overcrowding in ED. Salient examples using Shewhart-type control charts include the application of p-chart to monitor the variability of the number of patients leaving the ED [50], $\bar{x}$-chart to monitor the door-to-reperfusion time for patients who have acute ST myocardial infarction [51], and run

charts are developed to monitor the patient mortality rate [52] and daily demand in order to identify the start and end of the winter surge of pediatric patients in ED [53]. Unlike the Shewhart-type charts depended on only the current observation, the charts based on CUSUM and EWMA schemes accumulate information from past observations. For example, the authors in [54] implemented an EWMA chart to detect significant changes in the average number of deaths in the intensive care units of hospitals in Australia and New Zealand. The authors in [55] developed advanced CUSUM charts for monitoring the performance of typical queueing systems like ED with single queueing node. Kenyon et al. [56] applied X and S statistical process control chart to monitor the daily ED utilization of pediatric asthma emergency department. These methods focus on monitoring of specific quality indicators, such as the queue length of an individual queue, using univariate control charts. Service systems like the ED have a networked structure, so we cannot ignore the multidimensionality and granularity of the data obtained from electronic health records that can capture the delay experienced by patients at various stages of the care delivery process. To deal with that, Harrou et al. [57] proposed a principal component analysis (PCA)-based anomaly detection approach to monitor multiple correlated variables. They combined PCA modeling and the MCUSUM control chart to improve the accuracy in detecting the abnormal situations in ED.

1.2.3 Statistical Monitoring the Quality of Chronic Care Service

Unlike acute diseases that develop suddenly and last a short time, chronic diseases generally last a long period of time and require ongoing medical attention. Due to long-lasting damage to the body and brain, chronic diseases cannot be cured but only controlled. Thus, developing statistical process control method for monitoring the quality of chronic care service and treatment progress is critical to ensure the favorable treatment outcomes and prevent relapse. Examples of chronic medical conditions include HIV, asthma, diabetes, depression, substance use disorder, etc. Adeoti [58] applied CUSUM control scheme to detect changes

in the number of patients who tested positive to HIV/AIDS in Nigeria. Turner et al. [59] evaluated the use of different control charts or monitoring the lung function in asthma based on data from randomized trials. Hayati et al. [60] demonstrated Shewhart control chart is an effective, simple, and inexpensive method to identify occupational asthma. Dokouhaki and Noorossana [61] developed a control chart based on a two-state Markov model for monitoring the risk of epidemicity of diabetes based on auto-correlated discrete data. Aslam et al. [62] showed EWMA control chart with repetitive sampling is more effective than Shewhart control chart for detecting the shift in blood glucose levels in Type-II diabetes patients. Kaczmarek-Majer et al. [63] proposed a control chart based on weighted model averaging for monitoring the stability of patients with bipolar disorder. They claimed this method can take into account the uncertainty in the data and is more accurate and simple than other typical control charts. Cottrill et al. [64] applied p-chart to help identify adolescents and young adults with opioid use disorders, which allow early interventions to promote their initiation, engagement and retention in treatment.

First, medication non-adherence is a prevalent problem that affecting the quality of chronic care services. Proper medication adherence can decrease the occurrence rate of many major irrevocable health complications including death. In addition, poor medication adherence results in more than 100,000 mortalities each year in US and costs hundreds of billion dollars of healthcare spending annually [65]. Since improving medication adherence can achieve a significant benefit from both health and economic perspective, many methods have been used for the aim of monitoring medication adherence. Direct and indirect methods can be summarized in monitoring and measuring medication adherence. Direct methods of measuring the medication adherence include monitoring the drug concentration level in the patients' blood or urine. Indirect of measuring the medication adherence include self-reporting, pill-counting, analyzing patient's refill records, and measuring health outcomes, etc. And statistical monitoring techniques are considered as indirect monitoring methods. For example, Remien et al. [66] developed a backward stepwise regression analysis method

to determine which patient does not have a proper medication-taking behavior with the aim of improving Antiretroviral therapy adherence for HIV-infected patients. Hatoun et al. [67] proposed a Shewhart-type control chart by integrating with logistic regression models and negative binomial models to track the medication possession at discharge and medication fill rates after discharge for patients with asthma.

Furthermore, due to the long-lasting changes and damage in the brain and body caused by chronic diseases, relapse is common, so continuity of care or adherence to care is the critical factor for successful treatment for chronic diseases. However, in reality, most of the patients with chronic diseases have poor adherence to the duration of physician/therapy treatment and medication does or schedules. For example, the majority of the patients who are identified with AUD do not initiate treatment. And for those who initiated treatment, fewer than 15% continued in treatment [68]. Therefore, it is important to measure and monitor the treatment process for chronic diseases to identify care processes that lead to successful outcomes and patients whose adherence to care failed to occur in a timely manner and led to negative outcomes. Haberer et al. [69] developed a real-time approach for monitoring the antiretroviral therapy (ART) treatment adherence for HIV/AIDS patients. McHutchison et al. [70] applied χ^2 test and the 2-sided t test to assess the adherence to combination therapy with interferon or peginterferon plus ribavirin in chronic hepatitis C patients. Kubica et al. [71] adopted univariate and multivariate analysis for monitoring the adherence to treatment for patients with coronary artery disease (CAD) after myocardial infarction (MI) and identifying patients requiring personalized educational activities. In addition, underutilization of addiction treatment has been documented for over 20 years. Government reports claim that only 10% of people who need treatment receive it. Furthermore, retention in care whether the treatment is counseling or medication is poor with many people never getting past the initial assessment appointment. Recent efforts have focused on increasing capacity for treating opioid use disorder (OUD). For example, Matteliano et al. [72] developed a biopsychosocial-spiritual assessment model which is a comprehensive ap-

proach for monitoring and improving the adherence treatment of chronic opioid therapy for patients with persistent pain. Manchikanti et al. [73] proposed an evaluation tool including a chart review to monitor controlled substance intake for patients with chronic pain, which results in 50% reduction in opioid abuse. These researches did improve the access to care for patients with opioid use disorder, but not continuity of care or outcomes despite an increase in public and private expenditures. In the meantime, although access may have improved for the treatment of OUD, it has not increased for people with other substance use disorders. Thus new approach is needed to monitor the quality of system-level care in other substance use disorders treatment such as alcohol use disorder (AUD) treatment.

1.3 Overview and Organization of the Dissertation

This dissertation focuses on developing a series of statistical monitoring methods based on stochastic process models for improving the service quality in healthcare including acute and chronic care services. The detailed contributions and advancements of the developed statistical monitoring methods in each chapter as well as their applications are elaborated with details as follows.

As described in the previous section, effective monitoring of healthcare time-to-event data to improve the system quality has increasingly attracted attention from researchers in the area of statistical process control. However, many of the existing papers assume the healthcare time-to-event data as independent and identically distributed data or serially correlated discrete time stochastic processes. Very limited research has been conducted for monitoring continuous-time stochastic processes (CTSPs). To fill the gaps and to address the research need of monitoring continuous-time stochastic processes (CTSPs) based on time-to-event data. In Chapter 2, a novel statistical monitoring method is proposed for continuous-time stochastic processes with a focus on queueing processes under the emergency department(ED) setting. The proposed method is based on detecting a change in the intensity function of such processes, using an approximate likelihood ratio test. The

approximation method is both computationally easy for real-time implementation and well-suited for the introduction of penalization methods. Simulation results based on Markovian and non-Markovian queues show that the proposed methods can effectively detect temporal changes in the queueing process. A case study focusing on monitoring the waiting time of patients visiting an emergency department demonstrates the efficacy of the proposed method in a healthcare system.

The second type of continuous-time stochastic process faced in healthcare service system I want to address in this dissertation is the queueing networks. As described in the previous section, most of the existing statistical monitoring papers focus on monitoring single stage service process and they did not consider the multidimensionality in some complex service processes. While considering multiple stages in a service process with time-to-event data, one stage may affect the performance of next stage, thus the statistical monitoring methods focusing on individual stage are not appropriate to identify the abnormality for multiple-stage service processes. Further, we want to note that we observed most of the existing papers focus on monitoring the queue length or waiting time in a service system modeled as a queue, limited attention has been paid of detecting the changes of the system parameters like service rate, which is the key factor that reflect the service ability of a service system like ED. To fill these gaps, in Chapter 3, the statistical monitoring method is extended to accommodate complex service system structure with multiple nodes or stages. I proposed cumulative sum (CUSUM) control charts that monitor the queueing information collected in real-time from a queueing network (QN). We compare the proposed methods with existing statistical monitoring methods to demonstrate their ability to quickly detect a change in the service rate of one or more queues at the nodes in the QN. Simulation results show that the proposed CUSUM charts are more effective than existing statistical monitoring methods. The motivation for this research comes from the need to monitor the performance of a hospital emergency department (ED) with the goal of monitoring delays experienced by patients at various stages of the care delivery process in visiting the ED. A case study using

the data from the ED of a large academic medical center shows that proposed methods are a promising tool for monitoring the timeliness of care provided to patients visiting the ED.

In addition to monitoring the quality of acute care services such as emergency department, this dissertation also investigates quality monitoring methods for chronic care services, such as alcohol use disorder treatment processes. As described in the previous section, a Cascade of Care (COC) framework has been widely applied to improve system-level practice and treatment outcomes for various chronic medical conditions. However, very limited research has been conducted on monitoring the treatment process based on COC framework for alcohol use disorder (AUD). To fill this gap, in Chapter 4, the work aims to develop and test a model for measuring and monitoring the treatment processes of AUD using a COC framework. First, an innovative continuous-time stochastic process model is proposed to represent the dynamics of the COC for AUD treatment, from which benchmarks for COC can be developed by learning ideal patterns during different stages in care for AUD related to outcomes that indicate improved health. To the best of our knowledge, this study would be the first extension of the continuous-time stochastic modeling approach to AUD treatment processes. Then, a new statistical monitoring scheme is developed to identify the patients whose care deviated from the baseline model. Finally, key factor that is most correlated to undesirable health outcomes is identified, which would help clinicians develop subsequent interventions to promote treatment and improve outcomes for AUD.

The dissertation is organized as follows. Chapter 1 introduces the background and significance of monitoring the quality of healthcare services, and further presents a literature review on existing statistical monitoring methods for monitoring the quality of acute care service and chronic care service. Chapter 2 proposes a novel statistical monitoring method for detecting changes in the departure intensity function of ED service node by integrating with quadratic contrast estimation techniques. Chapter 3 proposes likelihood ratio based cumulative sum (CUSUM) control charts for monitoring the service rate of queueing network with time-inhomogenerous state dependent queues under the ED setting. Chapter 4

proposes continuous-time stochastic process model to measure and monitor the treatment process for patients with AUD based on the COC framework, from which benchmarks for COC can be developed and key factors that are correlated to undesirable health outcomes identified. In Chapter 5, the conclusion of this dissertation is drawn and the future research directions are discussed. Figure 1.2 gives an organizational diagram of this dissertation.

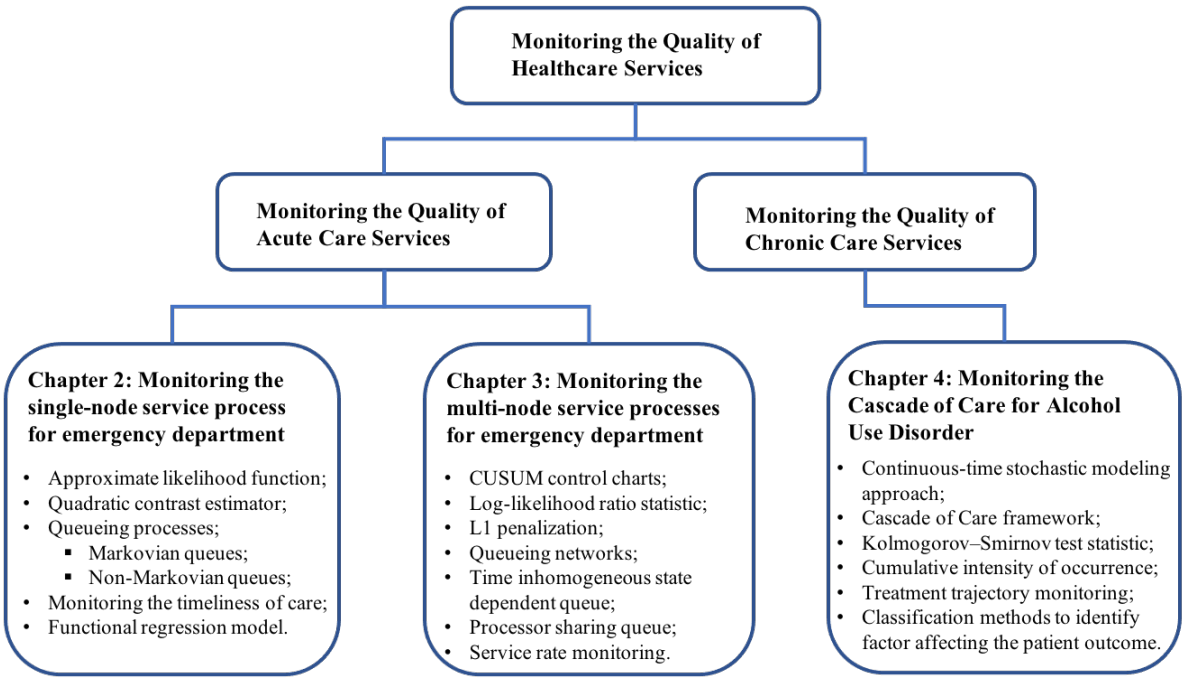


Figure 1.2: Organization of dissertation

Chapter 2: Monitoring Timeliness of Healthcare Delivery in Emergency Department Using Counting Processes

2.1 Overview

In recent years, effective monitoring of categorical and count data has increasingly attracted attention of researchers in the area of statistical process control. However, most of the existing research model categorical and count data streams as independent and identically distributed data or serially correlated discrete time stochastic processes. Very limited research has been conducted for monitoring continuous-time stochastic processes (CTSPs). This paper develops a novel statistical monitoring method for CTSPs with a focus on queueing processes. The proposed method is based on detecting a change in the intensity function of such processes, using an approximate likelihood ratio test. The approximation method is both computationally easy for real-time implementation and well-suited for the introduction of penalization methods. Simulation results based on Markovian and non-Markovian queues show that the proposed methods effectively detect temporal changes in the queueing process. A case study focusing on monitoring the waiting time of patients visiting an emergency department demonstrates the efficacy of the proposed methods in a healthcare system.

The methods researched in this paper can be used by operations managers in service enterprises, such as healthcare industries, for monitoring the timeliness of service provided to customers. The proposed method requires arrival and departure timestamps of customers from a queueing system when the system is considered ideal. This data is then used to define a metric for evaluating the queueing system's performance using real-time data. The proposed method does not require the arrival rate of the customers to be time-homogeneous. The experimental results show that the method is agnostic to classical Markov process as-

sumptions required in traditional performance modeling methods for queueing systems. The paper describes applying the proposed method to a timeliness-of-care monitoring problem in the emergency department of a large academic medical center. However, the method is expected to be broadly applicable to other service systems as well.

2.2 Introduction

Longitudinal and serially dependent categorical and count data occur in many engineering, financial, and biomedical processes. A few representative examples include – (1) study of manufacturing processes where the quality of each finished product is assigned a label, such as acceptable, repair, or discard; (2) the number of customers waiting in a queue at any given time-point of operation; and (3) utilization metrics of hospital emergency department (ED), which can be labeled as not busy, moderately busy, and highly busy during different times of the day. These examples illustrate applications where performance indicator metrics are categorical and count data. Therefore, monitoring categorical and count data stream is important to detect deterioration in the system from which the data is collected. Statistical process control (SPC) literature describe an extensive set of categorical and count data monitoring methods [74]. However, most of the existing research model categorical and count data streams as independent and identically distributed (i.i.d.) data or serially correlated discrete time stochastic processes. Very limited research has been conducted for monitoring continuous-time stochastic processes (CTSPs). To fill the research gap, this paper proposes a novel statistical monitoring method for CTSPs with a focus on queueing processes.

Queueing models have a wide range of application in stochastic modeling of service systems and contain vast amount of countable events such as the number of customers waiting in a queue. Unlike traditional statistical models used in SPC methods for count and categorical data that are based on probability distributions and transition probability distributions, queueing processes are often defined by intensity or rate functions. Applications that involve intensity functions include Markov processes, Poisson processes, Renewal processes and birth-

death processes where the intensities include the rate of arrival (or birth) and departure (or death) in queueing processes [75]. Figure 2.1 (a) and (b) are examples of time-homogeneous stochastic process with constant state transition intensity and time-inhomogeneous stochastic process with time-varying state transition intensity function, respectively.

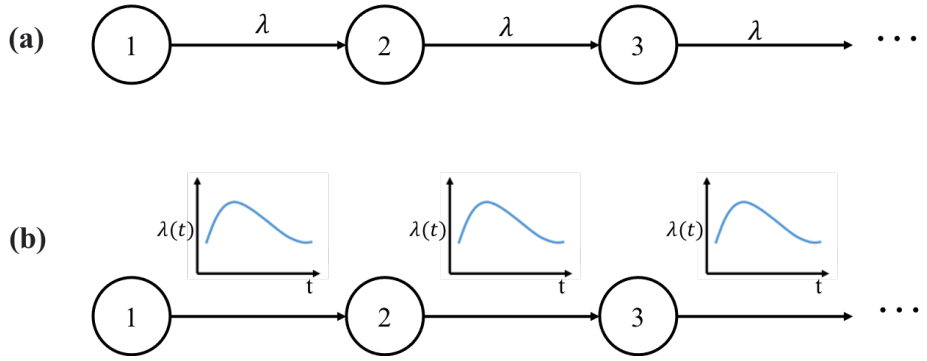
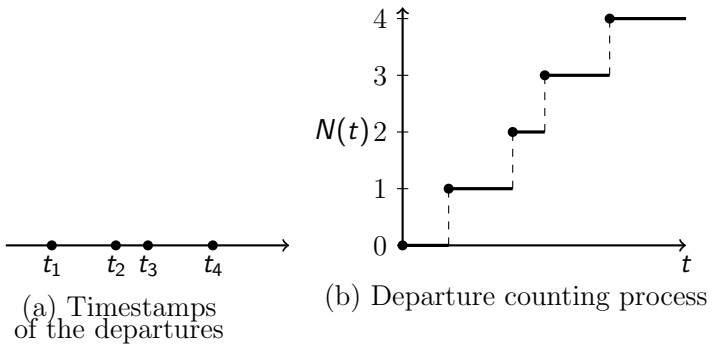


Figure 2.1: Illustration of stochastic processes: (a) Time-homogeneous stochastic process (b) Time-inhomogeneous stochastic process

The objective of this research is to develop a statistical monitoring scheme that detects changes in the rate of the departure process from a queueing system to evaluate, in real-time, whether the system performance has deteriorated or not. The departure process, which is defined as the real-time count of the number of customers processed in a queue (also considered as the number of departures from a queue), is modeled as a class of counting process that have predictable intensity. Figure 2.2 illustrates the procedure for transferring the observed departure timestamps to a counting process. This statistical monitoring scheme often involves two steps. The first step is the estimation of the intensity function from the data collected when the system is in control. The second step is the statistical test performed to label an observed departure timestamp sample as out of control if it deviates significantly from the in-control departure intensity function.

Maximizing the likelihood function is commonly used to estimate the departure intensity function and define a generalized likelihood ratio (GLR) test for detecting out-of-control samples of quality characteristics in various SPC methods. However, computing the maximization problem using this likelihood function of departure timestamps can be computa-



(a) Timestamps of the departures

(b) Departure counting process

Figure 2.2: Illustration of the proposed idea for monitoring the departures in a queue: (a) The timestamps of the departures (b) The departure intensity is monitored using a counting process that denotes the number of departures in [0,t], which is denoted as $N(t)$.

tionally prohibitive. In general, it does not have a general analytical solution and must be solved numerically. Additionally, a statistical monitoring statistic for real-time evaluation is meant to be computed for every time-point, which can further aggravate the computational burden. To overcome these issues, this paper proposes an approximate likelihood ratio test based on the quadratic contrast estimator for constructing the desired statistical monitoring scheme.

The proposed methods include a simple quadratic contrast test (SQCT) and a generalized quadratic contrast test (GQCT) that are analogous to the simple likelihood ratio (SLR) test and generalized likelihood ratio (GLR) test, respectively. Unlike GLR tests for counting processes based on MLE, the optimization step in GQCT has an analytical solution and does not require expensive numerical computations [76]. This is especially advantageous for computing a test statistic in real-time. Furthermore, the GQCT test results in a quadratic minimization problem and the results from Gaussian linear models can be used in computation and analysis of the test statistic. The proposed methods are compared with the traditional monitoring scheme using simulation study of single-server and multi-server Markovian and non-Markovian queueing systems. Further, the data from an ED is used to illustrate the application of the proposed method in the real world. The real case study is based on the problem of monitoring the waiting time of patients waiting to be assigned a bed in the ED of a large academic medical center.

The remainder of the chapter is organized as follows. Section 2.3 describes the counting process model for departure process and the quadratic contrast estimator for the intensity of the departure process. The statistical monitoring method is described in Section 2.4. A simulation study and a real case study are presented in Sections 2.5 and 2.6, respectively. Finally, the conclusions of this research and future research directions are described in Section 2.7.

2.3 Quadratic Contrast Estimation

This section describes the estimation of the intensity function from in-control data. For each in-control sample $j \in \mathcal{J}_0$, let $\tau_{i,j}$ denote the ith departure timestamp in sample j. The counting process $N_j(t)$ that counts the number of departures in $[0, t]$ is defined as:

$$N_j(t) = \sum_i \mathbb{I}(\tau_{i,j} \leq t).$$

It is assumed that $N_j(t)$ are non-explosive, i.e. the number of departures are almost surely finite over finite intervals. The intensity of the departure process is defined as follows

$$\mathbb{P}(N_j(t + dt) - N_j(t) = 1 | \mathcal{F}_t) = \lambda_j(t)dt, \tag{2.1}$$

where $\mathcal{F}_t$ denotes the information available until time t. For example, $\mathcal{F}_t$ could be the arrivals, departures, and number of servers observed until time t (In measure-theoretic discussion on stochastic processes, it would be referred to as the filtration to which the counting process is adapted.) The departure intensity $\lambda_j(t)$ in (2.1) is a predictable processes. That is, $\lambda_j(t)$ is not random given $\mathcal{F}_{t^-}$, where $\mathcal{F}_{t^-}$ denote the union of the sets $\mathcal{F}_s$ for $s < t$.

The proposed model is based on the assumption that the in-control queueing system will have an intensity processes:

$$\lambda_j(t) = \mu(t) D_j(t), \tag{2.2}$$

where $D_j(t)$ is a function of arrival and departure time stamps recorded before time t and $\mu(t)$ is common for all in-control queueing systems.

An example of a queueing system that consider the departure process from an $M_t/M_t/s_t$ queue, where the arrivals to the queue follow inhomogeneous Poisson process, customers spend an exponentially distributed time being served, and the number of servers changes over time. The departure intensity for an $M_t/M_t/s_t$ queue is

$$\lambda_j(t) = \mu(t) B_j(t), \tag{2.3}$$

where $1/\mu(t)$ is average service time at time t and $B_j(t)$ is the number of busy servers at time t. This is similar to a multiplicative intensity process commonly used in proportional hazard models [77].

Defining the statistical monitoring scheme requires estimating $\mu(t)$. The maximum likelihood estimation of $\mu(t)$ would involve maximization of:

$$\max_{\boldsymbol{\theta}} \prod_{j \in \mathcal{J}_0} \left[\left\{ \prod_{i=1}^{N_j(T)} \lambda_j(\tau_{i,j}) \right\} \exp\left(- \int_0^T \lambda_j(s) ds \right) \right], \tag{2.4}$$

where the maximization is with respect to a certain finite dimensional parameter $\boldsymbol{\theta}$ that defines $\mu(t)$. Typically, $\mu(t)$ is approximated as

$$\mu(t) = \sum_{k=1}^{K} \theta_k \phi_k(t), \tag{2.5}$$

for $t \in [0, T]$, where $\phi_1, \phi_2, \ldots \phi_K$ are a set of basis functions over the interval $[0, T]$ and $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots \theta_K]$ are the coefficients. In general, (2.4) does not have an analytical solution. But since it is equivalent to the following convex minimization problem:

$$\min_{\boldsymbol{\theta}} \sum_{j \in \mathcal{J}_0} \left[\int_0^T \lambda_j(s) ds - \sum_{i=1}^{N_j(T)} \log \lambda_j(\tau_{i,j}) \right], \tag{2.6}$$

and for which, numerical optimization methods work well for large sample sizes. However there are a few drawbacks of this estimation method. The functional form of the intensity function in (2.5) can be used to fit a large class of $\lambda_j(t)$ by increasing the number of basis functions K. For smaller in-control sample sizes, such as a service system that intends to use a few selected number of days as the in-control dataset, require additional model selection criteria to ensure overfitting is avoided. Such process typically includes adding prior knowledge about the departure process as constraints. Therefore an alternative to (2.6), which is referred to as the quadratic contrast estimator has recently been reported in literature [76].

The quadratic contrast that provides a simpler estimation method .

$$\min_{\boldsymbol{\theta}} \sum_{j \in \mathcal{J}_0} \left[\int_0^T \lambda_j^2(s) ds - \sum_{i=1}^{N_j(T)} 2\lambda_j(\tau_{i,j}) \right], \tag{2.7}$$

Such quadratic contrasts have recently received a lot of interest in high-dimensional statistical methods of counting processes [78, 79, 80, 81]. Also, this method has been used to detect multiple change points in counting processes [76]. There are some similarities to the Laplace approximation method used in Bayesian inference of count data. For $\lambda_j(t)$ defined as in (2.2) and (2.5), the minima $\hat{\boldsymbol{\theta}}$ for the minimization problem (2.7) is given as the solution to

$$\tilde{\mathbf{G}}\boldsymbol{\theta} = \tilde{\mathbf{n}} \tag{2.8}$$

where the kth row, lth column element of the matrix $\tilde{\mathbf{G}}$ is

$$\frac{1}{J_0}\sum_{j=\mathcal{J}_0}\int_0^T \phi_k(s)\phi_l(s)D_j^2(s)ds,$$

the $k$the element of $\tilde{\mathbf{n}}$ is

$$\frac{1}{J_0}\sum_{j=\mathcal{J}_0}\int_0^T \phi_k(s)D_j(s)dN_j(s) = \sum_{i=1}^{N_j(T)} \phi_k(\tau_{i,j})D_j(\tau_{i,j}),$$

and J_0 is the number of elements in set $\mathcal{J}_0$.

Departure processes with predictable intensities occur in a very large class of queues. In fact, even non-Markovian queues have a predictable departure intensity, where $\lambda(t)$ is a function of the number of customers being processed and the time they have spent in the process. The following proposition shows that the estimation method in (2.7) converges to the true departure estimated intensity if it is of the form defined by (2.3) and (2.5). We use the notation $[\mathbf{A}]_{ij}$ to denote the element in row i and column j of matrix $\mathbf{A}$.

Proposition 1. *Let $\pi(t) = \mathbb{E}(B^2(t))$, for $t \in [0, T]$. If the matrix $\mathbf{G}_0$, defined as*

$$[\mathbf{G}_0]_{kl} = \int_0^T \phi_k(s)\phi_l(s)\pi(s)ds,$$

is positive definite, then the solution to (2.8)

$$\hat{\boldsymbol{\theta}} = \tilde{\mathbf{G}}^{-1}\tilde{\mathbf{n}}$$

converges in probability to $\boldsymbol{\theta}$ that defined $\mu(t)$ in (2.5) as the size of the in-control sample size $|\mathcal{J}_0| \to \infty$.

Proof. Based on the theory of stochastic integral of counting processes,

$$\begin{aligned}\sum_{i=1}^{N_j(T)} \lambda(\tau_{i,j}) &= \int_0^T \lambda(s) dN_j(s) \\ &= \int_0^T \sum_{k=1}^{K} \theta_k \phi_k(s) B_j(s) dN_j(s) \\ &= \sum_{k=1}^{K} \theta_k \int_0^T \phi_k(s) B_j(s) dN_j(s) \\ &= \mathbf{n}_j^T \boldsymbol{\theta},\end{aligned}$$

where $\mathbf{n}_j = [n_{1,j}, n_{2,j}, \ldots n_{K,j}]$ and

$$\begin{aligned}n_{k,j} &= \int_0^T \phi_k(s) B_j(s) dN_j(s) \\ &= \sum_{i=1}^{N_j(T)} \phi_k(\tau_{i,j}) B_j(\tau_{i,j})\end{aligned}$$

Also,

$$\begin{aligned}&\int_0^T \lambda(s)^2 ds \\ &= \sum_{k=1}^{K} \sum_{l=1}^{K} \theta_k \theta_l \int_0^T \phi_k(s) \phi_l(s) B_j^2(s) ds \\ &= \boldsymbol{\theta} \mathbf{G}_j^T \boldsymbol{\theta},\end{aligned}$$

where the kth row lth column of matrix $\mathbf{G}_j$ is

$$g_{kl,j} = \int_0^T \phi_k(s) \phi_l(s) B_j^2(s) ds$$

Thus

$$\min_{\boldsymbol{\theta}} \sum_{j \in \mathcal{J}_0} \left[\int_0^T \lambda(s)^2 ds - \sum_{i=1}^{N_j(T)} 2\lambda(\tau_{i,j}) \right]$$
$$= \boldsymbol{\theta}^T \bar{\mathbf{G}} \boldsymbol{\theta} - 2\boldsymbol{\theta}^T \bar{\mathbf{n}}.$$

where

$$\bar{\mathbf{G}} = \frac{1}{J_0} \sum_{j \in \mathcal{J}_0} \mathbf{G}_j \quad \text{and} \quad \bar{\mathbf{n}} = \frac{1}{J_0} \sum_{j \in \mathcal{J}_0} \mathbf{n}_j$$

Thus the minimization problem in (2.7) is simplified as the solving for $\boldsymbol{\theta}$ that such that $\bar{\mathbf{G}}\boldsymbol{\theta} = \bar{\mathbf{n}}$ □

First we provide some of the definitions of the probability space on which the departure process N_j are defined. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space over which X is defined. Further, assume that $\mathcal{F}_t$ is an increasing, complete and right-continuous filtration such that $\mathcal{F}_T = \mathcal{F}$. The stochastic process N_j is adapted to $\mathcal{F}_t$ for all j.

Further, let

$$\ddot{N}(t) = \sum_{j=1}^{J_0} N_j(t),$$

and

$$\ddot{B}(t) = \sum_{j=1}^{J_0} B_j(t).$$

Since it is assumed that $\tau_{i,j}$, the transition times across all in-control sample j are different, $\ddot{N}(t)$ is a counting process adapted to the filtration $\mathcal{F}_t$, and has a Doob-Meyers decomposition given as

$$d\ddot{N}(t) = \lambda(t)dt + d\ddot{M}(t),$$

where $\ddot{M}(t)$ is a $\mathcal{F}_t$-martingale. Define

$$\frac{1}{J_0}U(\boldsymbol{\theta}) = \bar{\mathbf{G}}\boldsymbol{\theta} - \bar{\mathbf{n}}.$$

Since the estimating $\boldsymbol{\theta}$ involves solving $\frac{1}{J_0}U(\boldsymbol{\theta}) = \mathbf{0}$, convergence of $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}$ in probability is proved is the following conditions are satisfied:

1. $\frac{1}{J_0}U(\boldsymbol{\theta})$ converges in probability to $\mathbf{0}$ as $J_0 \to \infty$.
2. $\bar{\mathbf{G}}$ converges in probability to positive definite matrix $\mathbf{G}_0$.

The mentioned conditions are similar to the conditions that ensure the convergence in probability of the MLE for random variable. The quadratic contrast estimator satisfies these conditions and given in Lemma A.1 and A.2 prove that the two conditions are true for the estimator. Thus, when $\mathbf{G}_0$ is positive definite, the consistency of the quadratic contrast based estimate is proved.

Lemma A.1. $\frac{1}{J_0}U(\boldsymbol{\theta})$ *converges in probability to* $\mathbf{0}$ *as* $J_0 \to \infty$

Proof.

$$\begin{aligned}
U_k(\boldsymbol{\theta}) &= \sum_{l=1}^{K}\theta_l\int_0^T \phi_k(s)\phi_l(s)\ddot{B}^2(s)ds \\
&\quad - \int_0^T \phi_k(s)\ddot{B}(s)d\ddot{N}(s) \\
&= \sum_{l=1}^{K}\theta_l\int_0^T \phi_k(s)\phi_l(s)\ddot{B}^2(s)ds \\
&\quad - \int_0^T \phi_k(s)\ddot{B}(s)\{\lambda(s)ds + d\ddot{M}(s)\} \\
&= \sum_{l=1}^{K}\theta_l\int_0^T \phi_k(s)\phi_l(s)\ddot{B}^2(s)ds \\
&\quad - \sum_{l=1}^{K}\theta_l\int_0^T \phi_k(s)\phi_l(s)\ddot{B}^2(s)ds - \int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s).
\end{aligned}$$

Thus $U_k(\boldsymbol{\theta}) = -\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)$. Note that, as the stochastic integral with respect to a $\mathcal{F}_T$ martingale, $\mathbb{E}\left(\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)\right) = 0$.

In order to prove convergence in probability, following [82], we use the predictable variance process of the stochastic integral $\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)$ and apply Chebychev's inequality. Similarly, the variance of $\frac{1}{J_0}\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)$, denoted as $\mathbb{V}\left(\frac{1}{J_0}\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)\right)$, and

$$
\begin{aligned}
&\mathbb{V}\left(\frac{1}{J_0}\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)\right) \\
&= \frac{1}{J_0^2}\mathbb{E}\left(\int_0^T \phi_k^2(s)\ddot{B}^2(s)\lambda(s)ds\right) \\
&= \frac{1}{J_0}\mathbb{E}\left(\frac{1}{J_0}\int_0^T \sum_{l=1}^{K}\theta_l\phi_l(s)\phi_k^2(s)\ddot{B}^3(s)ds\right) \\
&= \frac{1}{J_0}\left(\int_0^T \sum_{l=1}^{K}\theta_l\phi_l(s)\phi_k^2(s)\mathbb{E}\left(\frac{1}{J_0}\ddot{B}^3(s)\right)ds\right)
\end{aligned}
$$

Using Chebychev's inequality,

$$
\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{J_0}\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)\right| \geq \epsilon\right) \\
&\leq \frac{1}{\epsilon^2}\mathbb{V}\left(\frac{1}{J_0}\int_0^T \phi_k(s)\ddot{B}(s)d\ddot{M}(s)\right) \to 0.
\end{aligned}
$$

Therefore, $\frac{1}{J_0}U(\boldsymbol{\theta})$ converges to 0 in probability. □

Lemma A.2. $\lim_{J_0\to\infty}\bar{\mathbf{G}}$ *converges in probability to* $\mathbf{G}_0$

Proof.

$$\begin{aligned}
&\lim_{J_0\to\infty} \frac{1}{J_0}\sum_{j=1}^{J_0} g_{kl,j} \\
&= \lim_{J_0\to\infty} \frac{1}{J_0}\int_0^T \phi_k(s)\phi_l(s)\sum_{j=1}^{J_0} B_j^2(s)ds \\
&= \int_0^T \lim_{J_0\to\infty} \frac{1}{J_0}\phi_k(s)\phi_l(s)\sum_{j=1}^{J_0} B_j^2(s)ds,
\end{aligned}$$

which follows from the Dominated Convergence Theorem, and

$$\begin{aligned}
&\int_0^T \phi_k(s)\phi_l(s)\left(\lim_{J_0\to\infty}\frac{1}{J_0}\sum_{j=1}^{J_0} B_j^2(s)\right)ds \\
&\to_{\mathbb{P}} \int_0^T \phi_k(s)\phi_l(s)\mathbb{E}(B_j^2(s))ds = \mathbf{G}_0
\end{aligned}$$

(by the Continuous Mapping Theorem)

□

In practice, the quadratic contrast estimation as given in (2.7) would over-fit the in-control data, which is particularly true if the number of basis function K is chosen to be significantly high or the data size is small. Therefore, a penalization term should be added. In the real-data based case study discussed later in the paper, we penalize $\boldsymbol{\theta}$ as follows:

$$\min_{\boldsymbol{\theta}} \sum_{j=1}^{J_0} \boldsymbol{\theta}^T\mathbf{G}_j\boldsymbol{\theta} - 2\mathbf{n}_j^T\boldsymbol{\theta} + \psi J_0\Lambda(\boldsymbol{\theta})$$

for a specified $\psi > 0$ and a chosen penalty function $\Lambda(\boldsymbol{\theta})$. The penalty function can be decided based on application. The quadratic and convex penalties are preferred for easier computation. Given the optimization problem is a quadratic minimization problem, several novel penalties such as total variation penalty and smoothness penalty can be easily added.

A cross-validation method can then be used to select the best value of ψ. The parameter estimation based on the in-control data is illustrated in Section 2.6.

2.4 Quadratic Contrast Tests

There are two popular likelihood ratio based SPC schemes – the simple likelihood ratio test and GLR test. If X denotes a test sample with likelihood $l(X, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the parameters of the distribution. A hypothesis test to test the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ vs. alternative hypothesis $H_1 : \boldsymbol{\theta} = \boldsymbol{\theta}_1$ is performed using the test statistic

$$T(X) = \frac{l(X, \boldsymbol{\theta}_1)}{l(X, \boldsymbol{\theta}_0)}.$$

When the above ratio $T(X) \geq c$ for a specified threshold c, the null hypothesis is rejected and it is not rejected otherwise. A statistical monitoring scheme based on such a test statistic would classify a sample X as out-of-control if $T(X) \geq c$, when $\boldsymbol{\theta}_0$ denotes the distribution parameters corresponding to the in-control system generating the samples. This test statistic is ideal when the out-of-control scenarios are approximated by the parameters $\boldsymbol{\theta}_1$. It requires the practitioners to specify $\boldsymbol{\theta}_1$ or define several $\boldsymbol{\theta}_1$ corresponding to many different out-of-control scenarios.

When stipulating the out-of-control distribution parameters is difficult, or when the out-of-control distribution lies in a parameter space Θ, a GLR test is often used. The test statistic for GLR test is

$$T(X) = \max_{\boldsymbol{\theta} \in \Theta} \frac{l(X, \boldsymbol{\theta})}{l(X, \boldsymbol{\theta}_0)}.$$

Recent, multivariate SPC methods have focused on penalization methods when $\boldsymbol{\theta}$ is multi-dimensional and the change in the parameters in the out-of-control distribution parameter occurs in only a few dimension. A penalized test GLR test statistic is given as

$$T(X) = \max_{\boldsymbol{\theta} \in \Theta} \frac{l(X, \boldsymbol{\theta})}{l(X, \boldsymbol{\theta}_0)} - \Lambda(\boldsymbol{\theta}),$$

where Λ a penalty function. Commonly used penalty function include the ℓ_2 and ℓ_1 or Lasso penalty.

For quality measures that follow multivariate normal distribution, $\Lambda(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||)$ is used. However, here we penalize $\boldsymbol{\theta}$ rather than $\boldsymbol{\theta} - \boldsymbol{\theta}_0$. The difference in the two approaches is that when $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ is penalized, the GLR test statistic is maximized likelihood for $\boldsymbol{\theta}$ that are not too far from $\boldsymbol{\theta}_0$. Whereas, when $\boldsymbol{\theta}$ is penalized, the GLR test statistic is maximized likelihood for $\boldsymbol{\theta}$ that are not too far from the origin of K-dimensional Euclidean space. We demonstrated that the second approach is suited for detecting decrease in intensity of Poisson process, which is discussed in Appendix B.

The procedure for detecting deterioration in system generating $N_j(t)$ is done by identifying the changes in the intensity parameter $\mu(t)$. The objective of this paper is to conduct a hypothesis test for any time in between $[0, T]$, which is given by the following hypothesis test:

$$\begin{aligned} H_0 &: \mu(s) = \mu_0(s) \ \forall s \in [0, t] \\ &\text{vs.} \\ H_1 &: \mu(s) \neq \mu_0(s) \ \forall s \in [0, t] \end{aligned} \tag{2.9}$$

for any $t \in [0, T]$, where $\mu_0(t)$ denotes the in-control intensity function. The unique aspect of the procedure developed in this paper, is that the hypothesis test in (2.9) is conducted for any $t \in [0, T]$.

The first test statistic is designed to perform the following hypothesis test:

$$\begin{aligned} H_0 &: \mu(s) = \mu_0(s) \ \forall s \in [0, t] \\ &\text{vs.} \\ H_1 &: \mu(s) = \mu_1(s) \ \forall s \in [0, t] \end{aligned} \tag{2.10}$$

for any $t \in [0, T]$, where $\mu_1(t)$ is assumed to be the out-of-control intensity function. This is a simple hypothesis test, and requires the knowledge of out-of-control intensity. Let

$$\mu_0(t) = \sum_{k=1}^{K} \theta_{0,k}\phi_k(t)$$

and

$$\mu_1(t) = \sum_{k=1}^{K} \theta_{1,k}\phi_k(t).$$

The proposed test statistic for testing the hypothesis in (2.10) is defined as:

$$\begin{aligned} \mathrm{SQCT}_j(t) = 2 \sum_{i=1}^{N_j(t)} \left(\mu_1(\tau_{i,j}) - \mu_0(\tau_{i,j})\right) D_j(\tau_{i,j}) \\ - \int_0^t \left(\mu_1^2(s) - \mu_0^2(s)\right) D_j^2(s) ds \end{aligned} \tag{2.11}$$

for any test CTMC X_j. Note that $\tau_{i,j} \leq t$ for $i \in \{1, 2, \ldots N_j(t)\}$. This test statistic is called the simple quadratic contrast test (SQCT). We introduce the following functions to simplify the calculation of SQCT:

$$g_{kl,j}(t) = \int_0^t \phi_k(s)\phi_l(s)D_j^2(s)ds$$

and

$$n_{k,j}(t) = \int_0^t \phi_k(s)D_j(s)dN_j(s)ds = \sum_{i=1}^{N_j(t)} \phi_k(\tau_{i,j})D_j(\tau_{i,j}).$$

It needs to be clarified that $n_{k,j}(T)$ and $g_{kl,j}(T)$ were previously denoted as $n_{k,j}$ and $g_{kl,j}$ respectively. Following previous convention, define the vector

$$\mathbf{n}_j(t) = [n_{j,1}(t)\ n_{j,2}(t)\ \cdots n_{j,K}(t)]^T$$

and $\mathbf{G}_j(t)$ is the matrix with the kth row and lth column element as $g_{kl,j}(t)$. Similarly, $\mathbf{G}_j$ and $\mathbf{n}_j$ have been used to denote $\mathbf{G}_j(T)$ and $\mathbf{n}_j(T)$ respectively. With this notation, the SQCT statistic is given as:

$$\begin{aligned}\mathrm{SQCT}_j(t) = 2\mathbf{n}_j^T(t)\left[\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0\right] \\ - \boldsymbol{\theta}_1^T\mathbf{G}_j(t)\boldsymbol{\theta}_1 + \boldsymbol{\theta}_0^T\mathbf{G}_j(t)\boldsymbol{\theta}_0,\end{aligned}$$

$\boldsymbol{\theta}_1 = [\theta_{1,1}\ \theta_{1,2}\ \cdots \theta_{1,K}]^T$ and $\boldsymbol{\theta}_0 = [\theta_{0,1}\ \theta_{0,2}\ \cdots \theta_{0,K}]^T$. This test statistic is similar to a simple likelihood ratio test for multivariate Gaussian random vectors.

For SPC problems where the out-of-control intensity functions are not known, we propose the generalized quadratic contrast test (GQCT) statistic as follows:

$$\begin{aligned}\mathrm{GQCT}_j(t) = \max_{\boldsymbol{\theta}} 2\sum_{i=1}^{N_j(t)}\left(\mu(\tau_{i,j}) - \mu_0(\tau_{i,j})\right)D_j(\tau_{i,j}) \\ - \int_0^t \left(\mu^2(s) - \mu_0^2(s)\right)D_j^2(s)ds - \Lambda(\boldsymbol{\theta}),\end{aligned} \tag{2.12}$$

where $\mu(t) = \sum_{k=1}^{K}\theta_k\phi_k(t)$, Λ is a penalty function. The idea of adding a penalty to the GQCT statistic comes form the success of penalized multivariate SPC methods [83]. The simulation studies reported in the paper use the elastic net penalty, and for which (2.12) simplifies as

$$\begin{aligned}\mathrm{GQCT}_j(t) = \max_{\boldsymbol{\theta}} 2\mathbf{n}_j^T(t)\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{G}_j(t)\boldsymbol{\theta} \\ - \left[2\mathbf{n}_j^T(t)\boldsymbol{\theta}_0 - \boldsymbol{\theta}_0^T\mathbf{G}_j(t)\boldsymbol{\theta}_0\right] - \eta||\boldsymbol{\theta}||_2^2 - \zeta||\boldsymbol{\theta}||_1,\end{aligned} \tag{2.13}$$

This optimization problem in (2.13) can be effectively solved using subgradient methods.

2.5 Simulation Study

This section reports the evaluation of the statistical power of the proposed method to detect changes in the service rate of Markovian and non-Markovian queues that experience

time-varying arrival rates and also have time-varying processing times. The basis functions used in this simulation study to approximate the in-control departure rate parameter $\mu(t)$ are:

$$\phi_1(t) = \frac{1}{2}$$
$$\phi_{2p}(t) = \cos(2p\pi t)$$
$$\phi_{2p+1}(t) = \sin(2p\pi t)$$

for $p = 1, 2, 3$, and 4. The simulations illustrate the application the proposed methods for detecting decrease and shift in departure rate of single-server and multi-server queues for both Markovian and non-Markovian queues. The departure intensity in the GQCT and SQCT models considers $\lambda_j(t) = \mu(t)B_j(t)$ for each test data j. This is the exact departure intensity for Markovian queues. However, the simulation results demonstrate that the performance of the monitoring schemes is robust to violation of the assumptions that service times are exponentially distributed.

We compare the SQCT scheme and elastic-net-penalized GQCT with a average length of stay (ALOS) monitoring scheme for detecting changes in departure rate, which has important application in monitoring service systems. It is a ALOS monitoring scheme and it is commonly used in queueing systems as a performance metric. The ALOS monitoring scheme is given as:

$$(A - \mathbb{E}(A|\boldsymbol{\theta}_0))^2.$$

where A is average time spent in the queue, which includes time in the queue and service time, of the customers departing the queue from $t = 0$ to $t = 1$.

In this simulation study, we use the type II error (also called miss detection rate) to measure the performance of the statistical monitoring methods. Type II error results from inferring that a process is in control when it is actually out of control.

2.5.1 Detecting Changes in Departure Rate of Single-server Queues

2.5.1.1 $M_t/M_t/1$ *Queue*

The queueing system studied here is a single-server queue to which the arrival of customers follows inhomogeneous Poisson process, has time-varying and exponentially distributed service time distribution, and processes customers on a first-come-first-serve basis. We set the arrivals intensity as:

$$10 - 5\sin(2\pi t),$$

for $0 \leq t \leq 1$, and the service times of customers who are processed starting at time t have an exponential distribution with rate

$$\frac{1.1}{\rho}\left(10 - 5\sin[2\pi(t-\omega)]\right),$$

When the queue is in-control, we set $\rho = 1$ and $\omega = 0$. The rational of selecting this queue is to simulate a service system that experience time varying arrival rates, and where the service rate has been designed to match the demand. When such a queue is out-of-control, the parameters ω and ρ will change. An increase in ρ implies that the service rate has become slower and a change in ω implies a temporal shift in the service rate function, which means that the peak service rate does not match the peak arrival, which is often the goal of optimized service rate in queues.

Figure 2.3 and 2.4 compares the described methods for detecting increase in ρ and change in ω. We can see both SQCT and GQCT are more sensitive than ALOS monitoring scheme

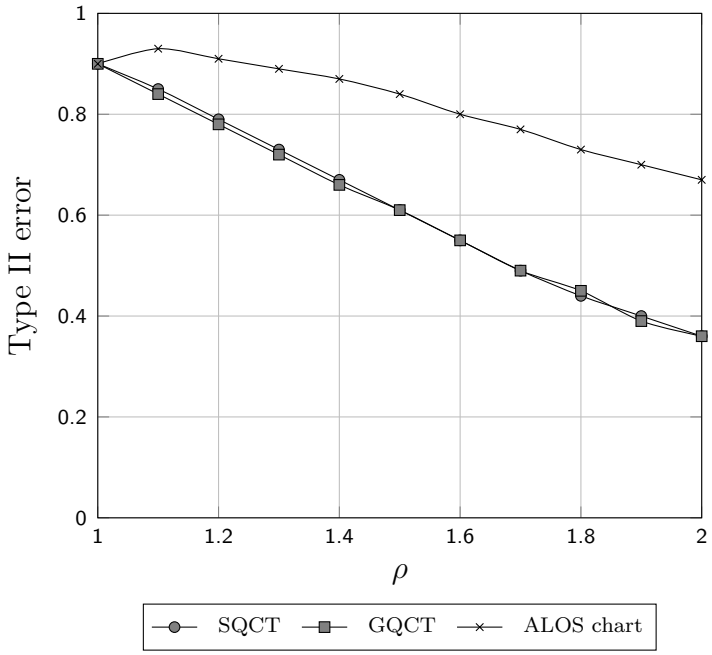


Figure 2.3: Detecting increase in ρ of a $M_t/M_t/1$ queue.

to detect the decrease and shift in departure rate of a $M_t/M_t/1$ queue, which demonstrates the benefit of using an approximate-likelihood-ratio-based approach of monitoring departure rate. In addition, SQCT and GQCT have similar performance in detecting the decrease in departure rate, while SQCT is better than GQCT in detecting the shift in departure rate. SQCT is more specific in detecting out-of-control scenarios but requires the out-of-control scenario to match the stipulated alternative. GQCT is more general in its application, but its performance is only slightly worse than SQCT.

2.5.1.2 $M_t/G_t/1$ *Queue*

The queue studied here is a single-server queue that arrivals are determined by an inhomogeneous Poisson process, and the service times have a time-varying general distribution.

We set the arrivals intensity as

$$10 - 5\sin(2\pi t),$$

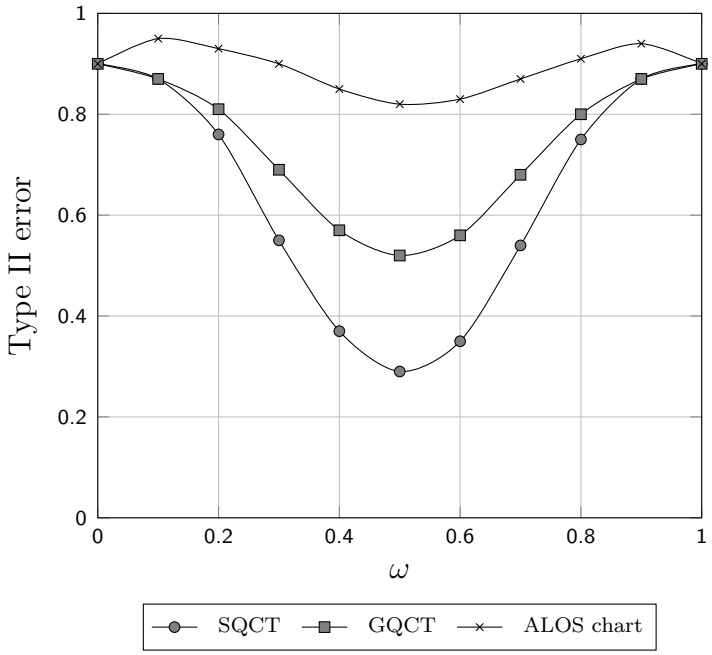


Figure 2.4: Detecting change in ω of a $M_t/M_t/1$ queue.

for $0 \leq t \leq 1$, and the service times of customers who are processed starting at time t are determined by a gamma distribution, which is denoted as

$$\text{Gamma}(\alpha, \gamma(t)),$$

where α is a shape parameter and $\gamma(t)$ is called a rate parameter. Here we set $\alpha = 4$ and

$$\gamma(t) = \frac{4.4}{\rho}\left(10 - 5\sin[2\pi(t-\omega)]\right),$$

for $0 \leq t \leq 1$. Based on the comparisons in Figure 2.5 and 2.6. We can see SQCT and GQCT are more sensitive for detecting the decrease and shift in departure rate than ALOS monitoring scheme. In addition, SQCT performs slightly better than GQCT for detecting changes in $M_t/G_t/1$ queue.

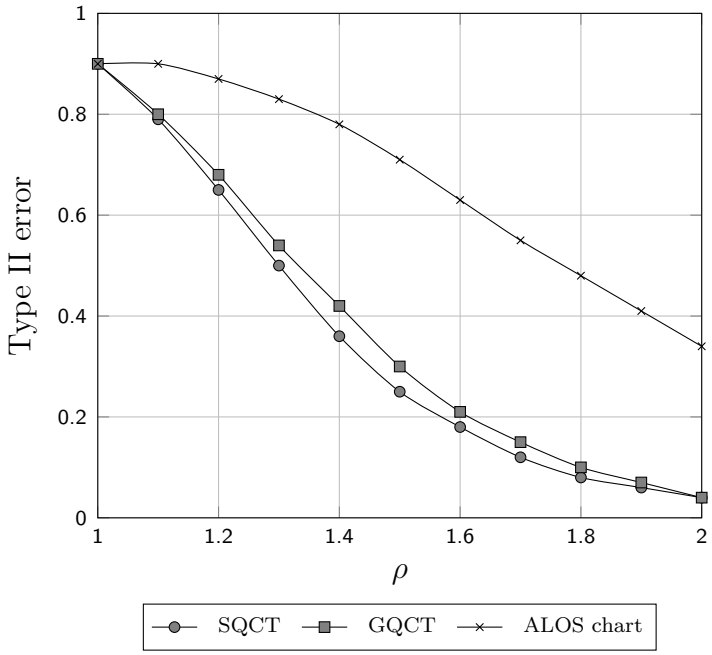


Figure 2.5: Detecting increase in ρ of a $M_t/G_t/1$ queue.

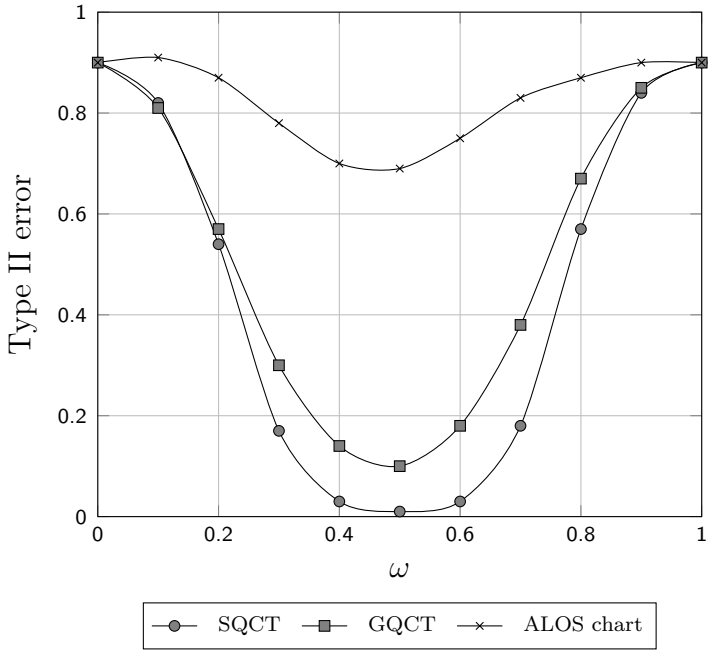


Figure 2.6: Detecting change in ω of a $M_t/G_t/1$ queue.

2.5.1.3 $G_t/G_t/1$ *Queue*

The queue studied here is a single-server queue that both inter-arrival times and service times have a time-varying general distribution. We set the inter-arrival times have a time-inhomogeneous gamma distribution, Gamma$(4, \gamma'(t))$, with

$$\gamma'(t) = 4/(10 - 5\sin(2\pi t)),$$

for $0 \leq t \leq 1$. Then, in order to match the demand, the service times of customers are determined by another time-varying gamma distribution, Gamma$(4, \gamma(t))$, with

$$\gamma(t) = \frac{4.4}{\rho}\left(10 - 5\sin[2\pi(t-\omega)]\right),$$

for $0 \leq t \leq 1$.

Based on the comparisons in Figure 2.7 and 2.8. We can see SQCT and GQCT are more sensitive for detecting the decrease and shift in departure rate than ALOS monitoring scheme. In addition, SQCT performs slightly better than GQCT for detecting changes in $M_t/G_t/1$ queue.

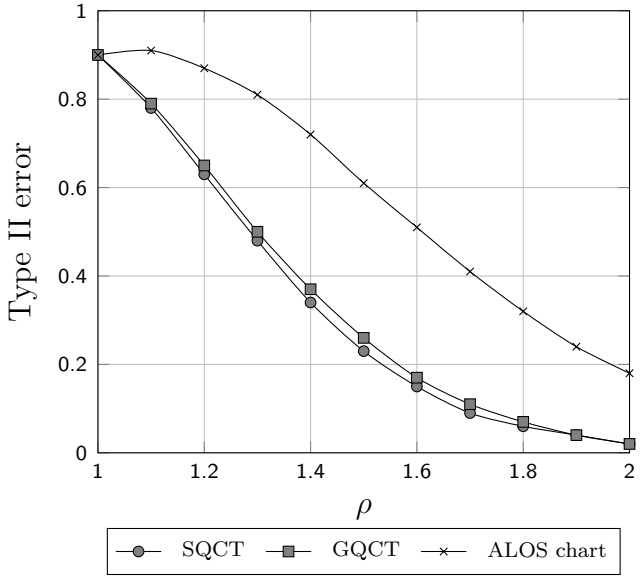


Figure 2.7: Detecting increase in ρ of a $G_t/G_t/1$ queue.

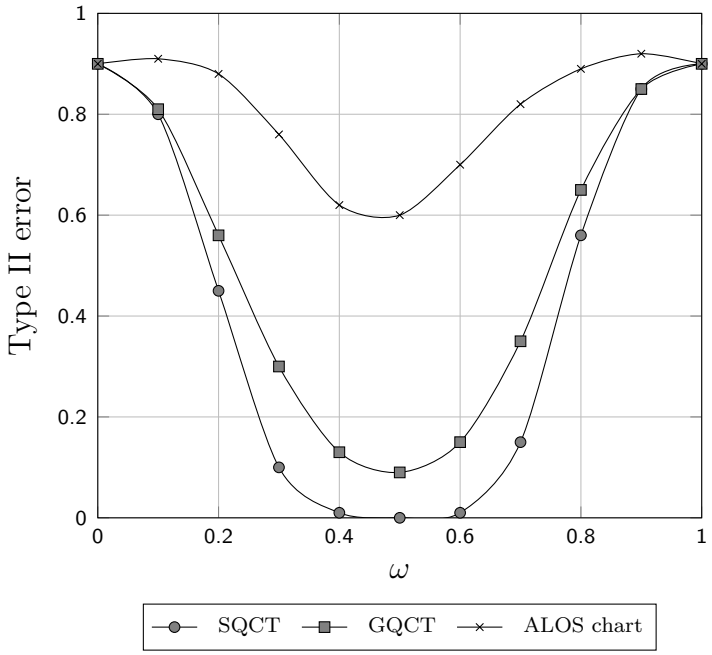


Figure 2.8: Detecting change in ω of a $G_t/G_t/1$ queue.

Sections 2.5.1.2 and 2.5.1.3 illustrate that users can use the SQCT and GQCT schemes even when the underlying assumption of exponential service times are violated.

2.5.2 Detecting Changes in Departure Rate of Multi-server Queues

In addition to the single server queues, the proposed methods can be extended to queues with multiple servers as well, such as $M_t/M_t/c$ queue, $M_t/G_t/c$ queue and $G_t/G_t/c$ queue. For $M_t/M_t/c$ queue, the arrivals are determined by an inhomogeneous Poisson process, has time-varying and exponentially distributed service time distribution, and processes customers on a first-come-first-serve basis. For $M_t/G_t/c$ queue, the arrivals are determined by an inhomogeneous Poisson process, and the service times have a time-varying general distribution. For $G_t/G_t/c$ queue, both the inter-arrival times and service times have a time-varying general distribution.

The simulation study results in Figure 2.9 - 2.14 show that SQCT and GQCT are quite sensitive for detecting the decrease and shift in departure rate of $M_t/M_t/5$ queue, $M_t/G_t/5$ queue and $G_t/G_t/5$ queue, while the ALOS monitoring scheme can barely detect any changes.

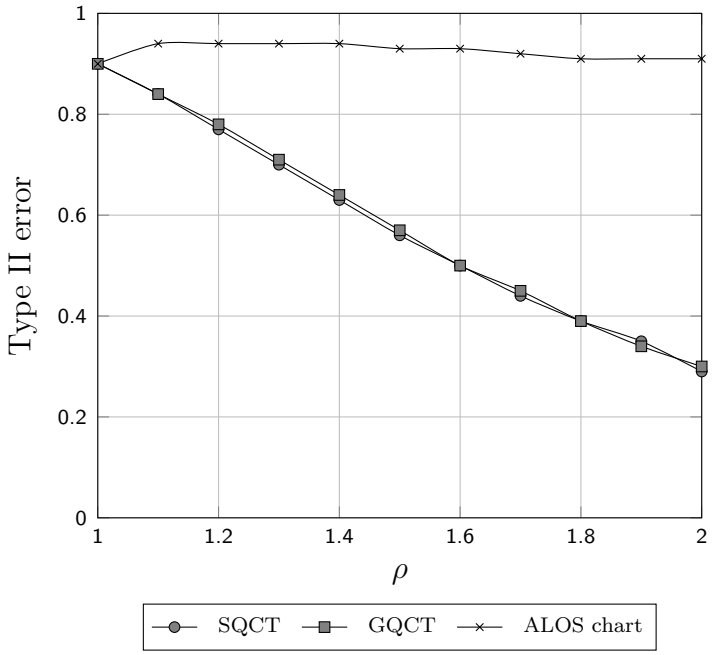


Figure 2.9: Detecting increase in ρ of a $M_t/M_t/5$ queue.

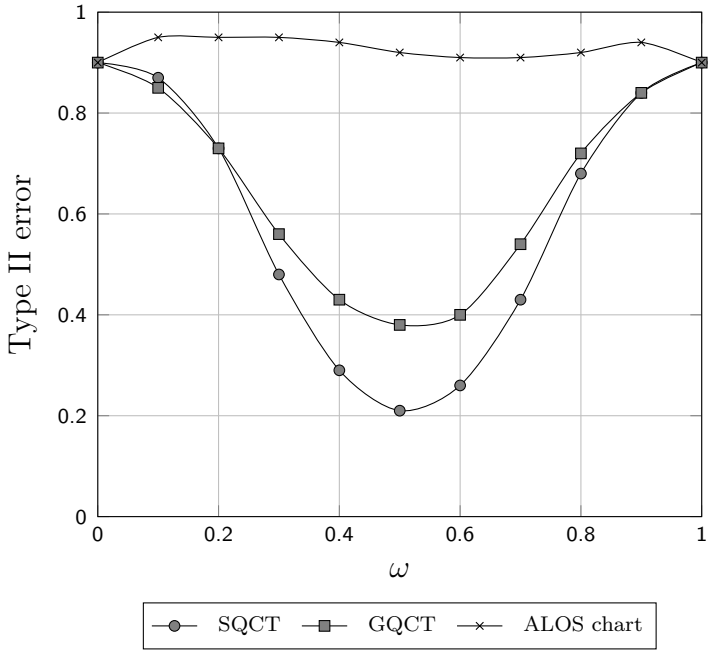


Figure 2.10: Detecting change in ω of a $M_t/M_t/5$ queue.

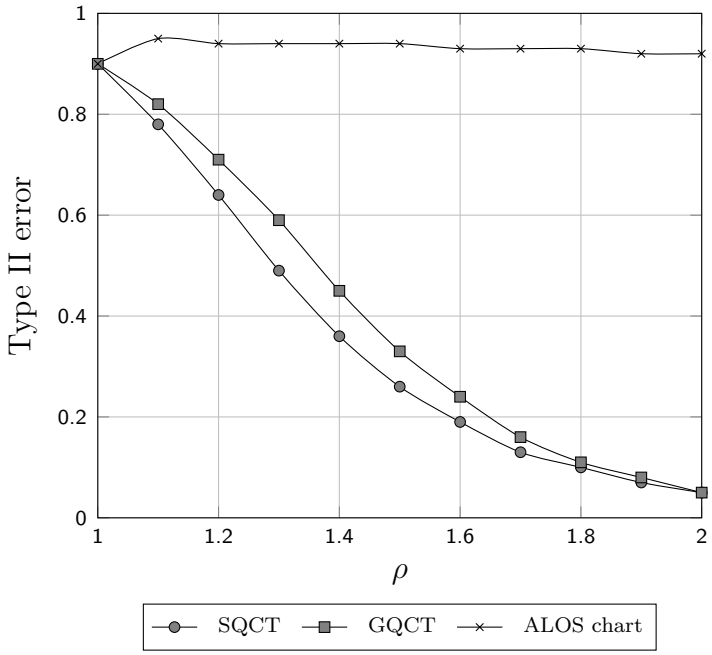


Figure 2.11: Detecting increase in ρ of a $M_t/G_t/5$ queue.

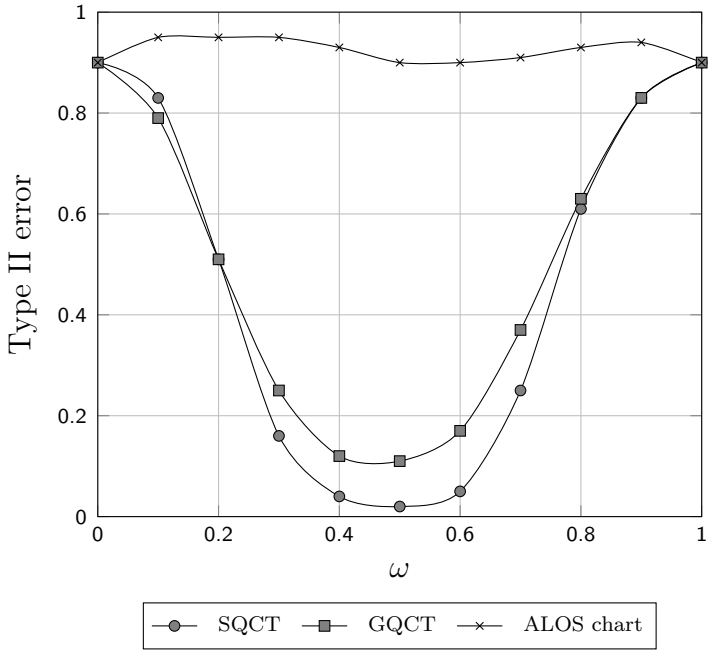


Figure 2.12: Detecting change in ω of a $M_t/G_t/5$ queue.

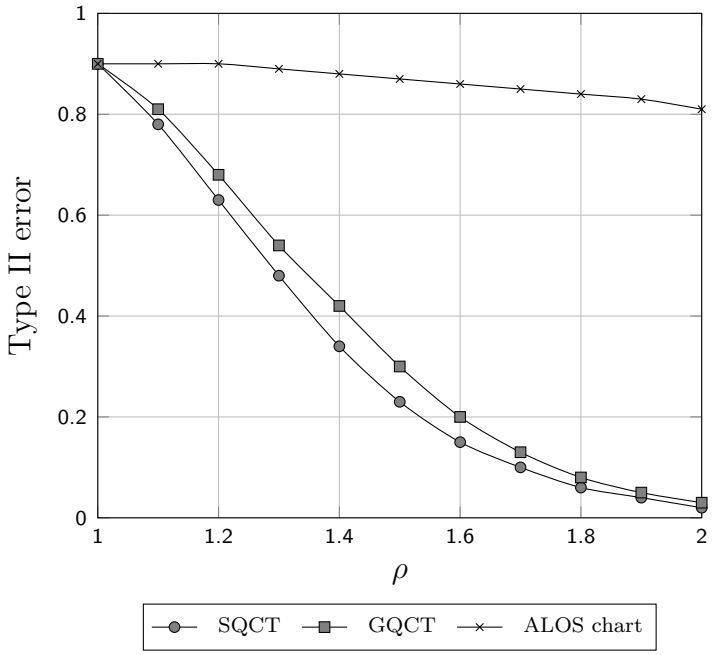


Figure 2.13: Detecting increase in ρ of a $G_t/G_t/5$ queue.

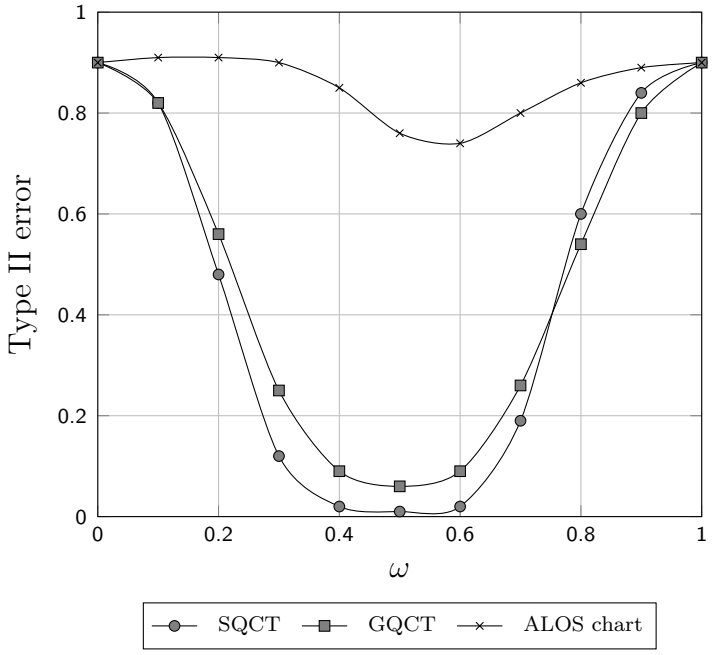


Figure 2.14: Detecting change in ω of a $G_t/G_t/5$ queue.

2.6 Monitoring the Waiting Patients Waiting Volume in an Hospital ED

In this section, our proposed monitoring schemes are tested to monitor the daily patient flow based on the emergency department (ED) patient visit data of a large academic medical center in the United States. Patients visiting the ED often have to wait for undesirably long time before they are assigned bed. Therefore, health services research has focused on various aspects of monitoring measures of delay in providing emergency care to patients. One such metric is the door-to-bed wait time [84]. We focus on monitoring the rate of bed assignment in the ED. The case-study here builds on the simulation study in Section 2.5. The arrival patients are considered to form a queue before they are assigned a bed, and thus monitoring the rate of bed assignment is equivalent to monitoring the departures from the arrival queue.

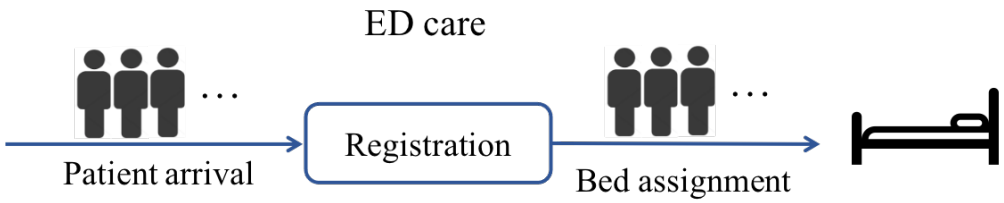


Figure 2.15: The ED patient flow from door to bed

We select the first 183 days in Year 2016 as training data, in which the days with patient average length of stay less than the 0.9 percentile (50.4 minutes) are used as in-control data. For parameter estimation of the in-control bed assignment rate, as described in 2.3, in order to avoid over-fitting cross-validation method is used to select the best value of the penalty, which we found is $\eta = 0.01$ and $\zeta = 100$. Figure 2.16 shows the in-control bed assignment rate of a day for the registration desk base on the in-control data, we can see that the registration desk has the lowest bed assignment rate at about 5AM and the highest bed assignment rate at about 11 AM.

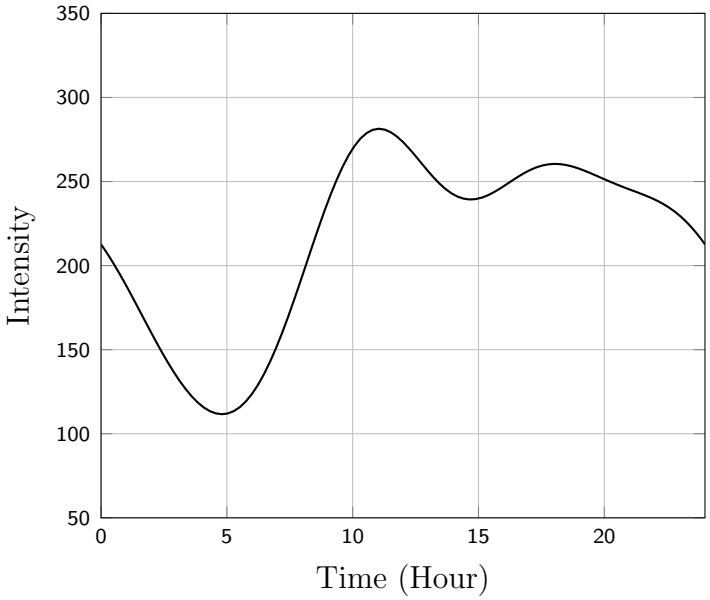


Figure 2.16: In-control bed assignment rate

We use the last 183 days in Year 2016 as test data set. We intend to use SQCT and GQCT to detect the decrease of the service rate. Then the ALOS chart designed to detect the increase of average door-to-bed wait time is used to compare with proposed SQCT and GQCT. The number of days classified as in-control and out-of-control are presented in the confusion matrix in Table 2.1 and 2.2. Table 2.1 shows that there are total 43 days in the test dataset are labeled as "out-of-control" by ALOS chart, while 35 days among the testing set are signed as "out-of-control" by SQCT, in which 16 days are identified as "out-of-control" by both SQCT and ALOS chart, and 19 days are identified as "out-of-control" by SQCT only. Similar results for the comparisons for GQCT and ALOS chart are given in Table 2.2.

Table 2.1: Confusion matrix for SQCT and ALOS chart

		SQCT		
	$N = 183$	Out-of-control	In-control	Total
ALOS chart	Out-of-control	16	27	43
	In-control	19	121	140
	Total	35	148	

In the test dataset, there are 12 days signaled out-of-control by both SQCT and GQCT but not the ALOS chart. To get further insight into the reason for this, study August 26,

Table 2.2: Confusion matrix for GQCT and ALOS chart

		GQCT		
	$N = 183$	Out-of-control	In-control	Total
ALOS chart	Out-of-control	16	27	43
	In-control	16	124	140
	Total	32	151	

2016 in further detail. Figure 2.17 shows the observed bed assignment rate and the in-control bed assignment rate. We can observe there is an overall decrease in the rate on August 26, particularly the peak hour rate. However, the mean and variance of the patient length of stay on August 26 were 32.77 minutes and 1.26 minutes respectively. This example illustrates a case where the average door-to-bed time did not deviate from the in-control average value, and, hence, the ALOS test statistic did not distinguish it as out-of-control. However, the intensity or rate of patients moving from waiting area to beds had clearly decreased on August 26, and we are able to detect it using the proposed methods.

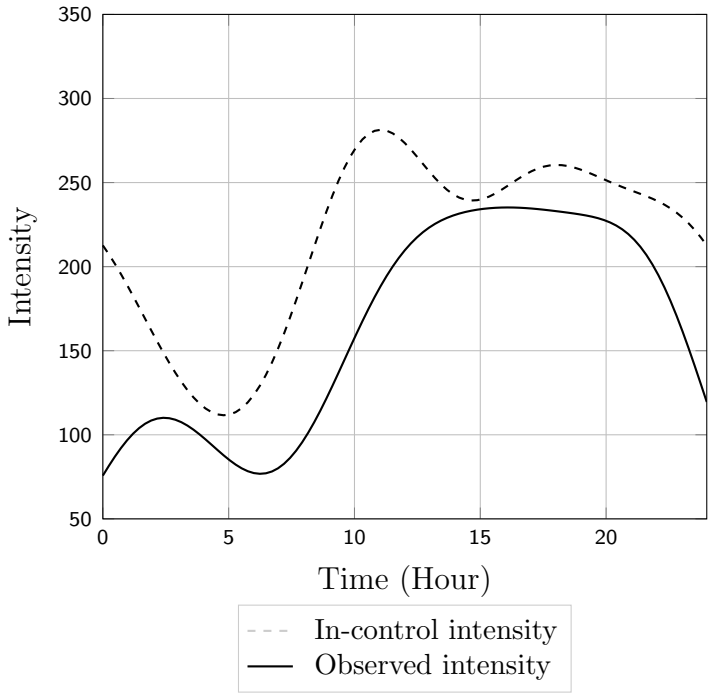


Figure 2.17: Intensity comparison on August 26, which was signaled out-of-control by both SQCT and GQCT but not the ALOS chart

2.7 Conclusions

In this paper, we have introduced a new statistical monitoring method for detecting changes in the departure intensity function of queues. The proposed method is based on an approximate likelihood function that alleviates the issue of needing to numerically maximize a complex likelihood function for estimating the in-control parameters and obtaining test statistics. There are two types of monitoring schemes that are proposed in the paper as alternatives to the SLR and GLR tests. Both methods were shown to detect changes in the intensity that is otherwise hard to detect using existing techniques. Besides, the proposed scheme can be used to identify changes in real-time, which is particularly complicated for inhomogeneous queueing systems. The efficacy of the methods is demonstrated by simulation studies and a real-data case study. The real-data case study analyzes the problem of monitoring the waiting time of patients visiting the ED of a major academic medical center.

There are several extensions of the method developed in this paper. Among them, an optimal detection scheme that minimizes delay in change detection for inhomogeneous CTSPs would have a wide range of applications. Further theoretical development of the GQCT method can lead to a better understanding of the type of penalty and the magnitude of the penalty that is ideal for detecting a specific kind of change in the intensity. Besides, monitoring the quality of care provided in the ED is an essential area of research in emergency medicine. We are working on developing other applications of the proposed methods to monitor the timeliness of care provided to patients visiting the ED. We anticipate that our ongoing research would result in several methodological developments and novel applications involving statistical monitoring of CTMCs.

Chapter 3: Statistical Monitoring of the Quality of Service in a Network of Queues with Application in Emergency Department

3.1 Overview

Queuing networks (QNs) are widely used stochastic models for service systems including healthcare systems, transportation systems, and computer networks. While existing literature has extensively focused on modeling and optimizing resource allocation in QNs, very little research has been done on developing systematic statistical monitoring methods for QNs. This paper proposes cumulative sum (CUSUM) control charts that monitor the queueing information collected in real-time from the QN. We compare the proposed methods with existing statistical monitoring methods to demonstrate their ability to quickly detect a change in the service rate of one or more queues at the nodes in the QN. Simulation results show that the proposed CUSUM charts are more effective than existing statistical monitoring methods. The motivation for this research comes from the need to monitor the performance of a hospital emergency department (ED) with the goal of monitoring delays experienced by patients at various stages of the care delivery process in visiting the ED. A case study using the data from the ED of a large academic medical center shows that proposed methods are a promising tool for monitoring the timeliness of care provided to patients visiting the ED modeled as a QN.

3.2 Introduction

A QN is the representation of a service system consisting of a network of servers. Each node of a QN consists of a set of servers processing or serving arriving entities, such as

customers in a service system or packets in a computer network [85]. In recent years, QNs have been widely used in modeling many service systems, such as manufacturing systems [86], computer networks [87], transportation systems [88], and healthcare systems [89, 90, 91]. Especially in healthcare, QN models have been found valuable in modeling the flow of patients in the hospital emergency departments (ED).

There is a rich body of research on resource allocation and decision-making in ED using QN models. For example, Cochran and Roche developed an open QN model to increase the capacity of an ED for patient care by considering various types of arrival patterns and volumes in patients [92]. Vass and Szabo applied QN models to determine the optimal allocation of trained personnel and specialized equipment in ED [93]. Huang et al. [94] and Xie et al. [95] presented QN models that consider triage patients as a multi-class queuing system to control the priority of patients' treatments. They model the ED as a traditional queuing system, such as M/M/1 and M/M/k queues, where the service capacity is bounded and constant. However, the available medical staff and resources in ED is time-varying in reality, which make traditional queuing systems with fixed service capacity a poor fit for ED. Therefore, Shi et al. [96] adopted a processor-sharing queue, where the service rates are functions of the number of patients over time, to study how to effectively integrate a new diagnostic test into the clinical environment in ED. They demonstrate the processor-sharing queue with a state-dependent service rate function is more flexible to accommodate the complexities commonly seen in the ED environment compared to the traditional queuing setting.

Resource allocation using performance modeling tools like QNs addresses the problem of optimizing scare emergency medical resources. But such problems are typically part of long-term operational decision making in the ED. For example, changing the staffing schedules too frequently could be opposed by ED healthcare providers [84]. Therefore, it is important to augment such performance modeling methods with real-time performance monitoring methods, which will ensure the adherence to a high quality of care and detect deterioration

in the hypothesized optimal flow of patients. Statistical process control (SPC) charts are increasingly being used in healthcare to monitor and measure the process variation and identify changes that indicate deterioration in quality [97]. It is important to note that the Center for Medicare and Medicaid Services requires that the hospitals report performance measures of the EDs, such as average length of stay of patients visiting the ED. Deterioration of these indicators can quickly bring down the quality of care. Research shows that lower service rates in ED can result in longer queue length and waiting time, which might increase the risk of adverse outcomes for patients [98, 99]. Therefore, it is imperative to develop statistical performance monitoring methods for evaluating the quality of healthcare delivery in the ED.

Statistical process control (SPC) methods have been studied in the context of monitoring the quality of service in the ED [48]. Salient examples using Shewhart-type control charts include the application of p-chart to monitor the variability of the number of patients leaving the ED [50], $\bar{x}$-chart to monitor the door-to-reperfusion time for patients who have acute ST myocardial infarction [51], and run charts are developed to monitor the patient mortality rate [52] and daily demand in order to identify the start and end of the winter surge of pediatric patients in ED [53]. Unlike the Shewhart-type charts depended on only the current observation, the charts based on CUSUM and EWMA schemes accumulate information from past observations. For example, the authors in [54] implemented an EWMA chart to detect significant changes in the average number of deaths in the intensive care units of hospitals in Australia and New Zealand. The authors in [55] developed advanced CUSUM charts for monitoring the performance of typical queuing systems with single queuing node. These methods focus on monitoring of specific quality indicators, such as the queue length of an individual queue, using univariate control charts. Service systems like the ED have a networked structure, so we cannot ignore the multidimensionality and granularity of the data obtained from electronic health records that can capture the delay experienced by patients at various stages of the care delivery process. Therefore, a multivariate statistical

monitoring scheme based on advanced stochastic models like QNs is crucial and needs to be developed.

The most appropriate and widely used multivariate control charts are multivariate EWMA (MEWMA) and multivariate CUSUM (MCUSUM) charts. Their good performance in monitoring the changes of process means, especially for small changes, has been validated by many papers [100]. However, these methods assume the process data follow a time-homogeneous multivariate normal distribution. It needs to be clarified that, many large sample approximations of queue performance metrics, such as diffusion approximation, also follow multivariate normal distribution [101]. In practice, the normality assumption is usually difficult to justify for a real time queuing performance metric obtained from a nonstationary QN, so that the statistical properties of MEWMA and MCUSUM charts could be affected. In addition, we observed most of the existing papers focus on monitoring the queue length or waiting time in a service system modeled as a queue, limited attention has been paid of detecting the changes of the system parameters like service rate, which is the key factor that reflects the service ability of a service system like ED.

To overcome the limitation caused by the multivariate normal distribution assumption and fill the gap of monitoring the system parameters of a service system, this paper proposes new CUSUM charts based on the likelihood ratio statistics to monitor the service rates for a QN with time-inhomogeneous state-dependent queues. The likelihood ratio statistics pose no constraint to the underlying process distribution and have been demonstrated to be generally more powerful than other alternative methods [102]. The proposed methods are evaluated based on the delay in detecting the change in service rate of one or more nodes of the QN. Our simulation results show that the proposed charts are more effective compared with conventional MCUSUM and MEWMA charts. Also, a real case study focusing on monitoring the daily patient flow of an emergency department demonstrates the efficacy of the proposed methods in real application.

The remainder of the chapter is organized as follows. Section 3.3 introduces the QN model, and Section 3.4 introduces the statistical monitoring scheme for the QN. Section 3.5 derives the CUSUM charts based on different likelihood ratio statistics to monitor the service rate of QN. Their numerical performances are investigated in Section 3.6. In Section 3.7, we demonstrate the application of the proposed methods using a real-data example from the ED of a large academic medical center. Finally, the conclusions of this research and future research directions are described in Section 3.8.

3.3 Queuing Network Model

A QN is a network of queues with queues at every node of the network. Entities (e.g., customers, patients, and data packets) processed in the QN arrive at a node in the network from either outside the network or a different node in the network and upon completion of processing at the node the entities move to a different node or leave the network. The structure of the network, the arrival process to the network, and the service policy are used to classify the type of QN. In this paper, we shall focus on studying the open Jackson network, which is a very commonly used QN in practical applications of queuing theory in service systems [103]. In an open Jackson network, external arrival can happen to any of the nodes and customers can leave the system from any of the nodes. A customer completing service at one node in the open Jackson network can either move to another queue node with some probability or leave the system.

Consider an open network with I nodes where the service rate of the nodes depends on the number of customers at each node. In this network, external arrivals to node i, for $i \in \{1, 2, \cdots, I\}$, occurs as a Poisson process with rate $\lambda_i(t)$, where t denotes the time of a day. Let $B_i(t)$ denote the number of customers at node i at time t. We assume that the service rate of queue at node i follows an exponential distribution with rate $\mu_i(t) = f_i(B_i(t), \theta_i)$, where θ_i is the vector of parameters that define f_i. The arrival process and service time

for each node are assumed to be mutually independent. The data from a queue network consisting of queues indexed by $i = 1, 2, \ldots I$ will have following events:

1. $\tau_i^1, \tau_i^2, \ldots \tau_i^{A_i(t)}$: The external arrivals to a queue node between times $[0, t]$ are independent of everything happening inside the network.

2. $\delta_i^1, \delta_i^2, \ldots \delta_i^{D_i(t)}$: The departures from a queue node between times $[0, t]$ are independent of everything except the number of customers at $\delta_i^1, \delta_i^2, \ldots \delta_i^{D_i(t)}$.

3. $e_i^1, e_i^2, \ldots e_i^{D_i(t)}$: The index of the queue that a customer leaving node i joins. $e_i^{n_i} = 0$ indicates that the n_ith customer leaving node i is either deterministic or dependent on the transition probabilities p_{ij}, where p_{ij} is the probability of a customer leaving node i to join node j and $p_{i0} = 1 - \sum_{j=1, j\neq i}^{I} p_{ij}$.

Since,

$$\mathbb{P}\left(t \leq \delta_i^{n_i} < t + dt | \,|\, \text{all events that have occurred on or before } t\right) = \mu_i(t)dt$$

for $n_i \in 1, 2, \ldots N_i(t)$ and

$$\mathbb{P}\left(t \leq \tau_i^{a_i} < t + dt | \,|\, \text{all events that have occurred on or before } t\right) = \lambda_i(t)dt$$

for $a_i \in 1, 2, \ldots A_i(t)$. The log likelihood of this data is given as

$$l(t, \Theta) = \sum_{i=1}^{I} \sum_{n_i=1}^{D_i(t)} \log \mu_i(\delta_i^{n_i}) - \int_0^t \mu_i(s)ds + \sum_{i=1}^{I} \sum_{a_i=1}^{A_i(t)} \log \lambda_i(\tau_i^{a_i}) - \int_0^t \lambda_i(s)ds + \sum_{i=1}^{I} \sum_{n_i=1}^{D_i(t)} \frac{\mathbb{I}(e_j^{n_i} = j)}{D_i(t)} \log p_{ij}.$$

Here $\Theta = [\theta_i]$, vector obtained from concatenating the service rate parameters from node i for $i = 1, \ldots I$. Let t_n denote the nth event in the ordered list of all arrivals ($\tau_i^{n_i}$) and departures ($\delta_i^{n_i}$), and let $l_n(\Theta) = l(t_n, \Theta)$. Thus,

Then, the log-likelihood function for the observed sample path in $(0, t_n]$ becomes

$$\begin{aligned} l_n(\Theta) = &\sum_{i=1}^{I} \sum_{n_i=1}^{D_i(t_n)} \log \mu_i(\delta_i^{n_i}) - \int_0^{t_n} \mu_i(s)ds + \\ &\sum_{i=1}^{I} \sum_{a_i=1}^{A_i(t_n)} \log \lambda_i(\tau_i^{a_i}) - \int_0^{t_n} \lambda_i(s)ds + \\ &\sum_{i=1}^{I} \sum_{n_i=1}^{D_i(t_n)} \frac{\mathbb{I}(e_j^{n_i} = j)}{D_i(t_n)} \log p_{ij}. \end{aligned} \tag{3.1}$$

In the following sections, instead of the likelihood ratio function, our proposed CUSUM charts are designed based on the log-likelihood function (3.1), which is both computationally easy and well-suited for the introduction of penalization methods.

3.4 Statistical Monitoring Scheme for Queueing Network

The likelihood function in (3.1) can be used to monitor any change in the QN. However, in practice, detecting deterioration in the performance of a QN caused by one or more queues at the nodes of the QN slowing down is more relevant than other types of changes. This detection problem is also the primary focus of the related research reviewed in Section 1. Therefore, we focus on building a statistical monitoring scheme to detect the change in the service rate of a QN. Let $\Theta_0 = \{\theta_1^0, \theta_2^0, \cdots, \theta_I^0\}$ represent the parameters related to service rate of each node in the QN when the system is in control, referred to as the in-control parameter, and let the parameter $\Theta = \{\theta_1, \theta_2, \cdots, \theta_I\}$ denote the parameters related to true service rate of each node. Hence, if the system is in control, then $\Theta_0 = \Theta$. The statistical monitoring scheme in this paper focuses on monitoring the QN only when an event such as arrival, departure, or movement of an entity from one node to another occurs. It is a framework also considered in prior research on monitoring single server queues [104]. Therefore, the monitoring statistic is only updated at t_n, which represents the time when n^{th} event occurs.

Therefore, for each t_n, a test statistic h_n is defined to test the following hypothesis

$$H_0 : \Theta = \Theta_0$$

vs.

$$H_1 : \Theta \neq \Theta_0.$$

A decision rule is defined to test this hypothesis in a CUSUM scheme as follows:

$$h_n > g$$

where g is the threshold value. Once the CUSUM statistic h_n exceeds the control limit g, an alarm is triggered. A generated alarm means that the observed process is classified as out of control. Then the time t_n where such an out-of-control signal first happens is used to define the run length n. Here, the control limit g is determined such that the average run length (ARL) under the in-control scenario, denoted by ARL_0, meets the specified value [105]. The CUSUM statistic is defined in the next section.

3.5 Proposed CUSUM Charts

The CUSUM chart, introduced by [106], is one of the most popular sequential change-detection methods used in statistical quality control. It is based on not only current observations but also past observations. It has been demonstrated that the conventional CUSUM chart and its modifications are very effective in detecting a large class of change in model parameters [55, 107, 108]. Therefore, we develop CUSUM charts for monitoring the deterioration in the service rate of queues in a QN. In this section, CUSUM charts that are based on the likelihood ratio statistics are proposed to monitor the service rate of QNs. The first is a simple CUSUM (SCUSUM) chart, which is best suited when the practitioners can specify the potential out-of-control parameters. However, the performance of the SCUSUM chart might deteriorate if the real out-of-control parameters were far from the hypothesized out-of-

control parameters. So, the general likelihood ratio and penalized likelihood ratio, which are computed by maximizing the likelihood ratio and penalized likelihood ratio respectively, are developed to construct the second type of CUSUM charts. They are called the generalized CUSUM (G-CUSUM) chart and penalized CUSUM (P-CUSUM) chart.

3.5.1 The SCUSUM Chart

For deriving the SCUSUM chart based on the likelihood ratio, a specified out of control parameter is needed. Let

$$\Theta_1 = \{\theta_1, \theta_2, \cdots, \theta_I\} = \{(1+\Delta_1)\theta_1^0, (1+\Delta_2)\theta_2^0, \cdots, (1+\Delta_I)\theta_I^0\}$$

represent the specified out of control service rates. Where Δ_i denotes the hypothesized degree of a shift away from in control parameter in μ_i^0. The sign of Δ_i should be consistent with the actual change direction of the service rate for each node $i = \{1, 2, ..., I\}$. For instance, if we are interested in detecting the decrease of the service rate for node i, then δ_i should be set as a negative value such as -10%.

Then, based on equation (3.1), we denote the log-likelihood ratio after the n^{th} event under the observed complete sample path $\{X(t_n)\}$ as ξ_n, which is

$$\begin{aligned} \xi_n &= l_n(\Theta_1) - l_n(\Theta_0) \\ &= \left[\sum_{i=1}^{I} \sum_{n_i=1}^{D_i(t_n)} \log \frac{f_i(B_i(\delta_i^{n_i}), \Theta_1}{f_i(B_i(\delta_i^{n_i}), \Theta_0)} - \int_0^{t_n} \left(f_i(B_i(s), \Theta_1) - f_i(B_i(s), \Theta_0)\right) ds \right] \end{aligned} \tag{3.2}$$

Thus, the SCUSUM statistic h_n^s is defined as

$$h_n^s = \max\{0, h_{n-1}^s + \xi_n - \xi_{n-1}\} \qquad \text{where} \quad h_0^s = 0.$$

3.5.2 The G-CUSUM and P-CUSUM Charts

The SCUSUM chart requires a set of specified design parameters, like Δ_i that indicate the type of change that the users want to detect. However, in practice, it is difficult for users to know the potential change in advance. For example, the user may need to specify the specific nodes of a QN that have the potential to slow down. For such cases, the SCUSUM chart may perform poorly when the actual change is different from the assumed change. As a solution to this problem, the specified service rate Θ_1 can be replaced by the maximum likelihood estimate (MLE) of the service rate for the server in each node by maximizing the log-likelihood ratio. The resulting CUSUM scheme results in a generalized-likelihood-ratio-based G-CUSUM chart.

Let ξ_n^g denote the generalized log-likelihood ratio after the n^{th} event, which is the maximum of the log-likelihood ratio in (3.2), that is

$$\begin{aligned}\xi_n^g &= \max_{\Theta}\left[\sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t_n)}\log\frac{f_i(B_i(\delta_i^{n_i}),\Theta)}{f_i(B_i(\delta_i^{n_i}),\Theta_0)}-\int_0^{t_n}\left(f_i(B_i(s),\Theta)-f_i(B_i(s),\Theta_0)\right)ds\right], \\ &= \sum_{i=1}^{I}\xi_{n,i}^g\end{aligned} \tag{3.3}$$

where

$$\xi_{n,i}^g = \max_{\Theta}\log\frac{f_i(B_i(\delta_i^{n_i}),\Theta)}{f_i(B_i(\delta_i^{n_i}),\Theta_0)}-\int_0^{t_n}\left(f_i(B_i(s),\Theta)-f_i(B_i(s),\Theta_0)\right)ds.$$

This maximization problem is simplified when f_i is a linear function in θ_i. Assume that f_i is functional linear model where:

$$f_i(B_i(t),\theta_i)) = \phi_i(B_i(t))^T\theta_i$$

where $\phi_i(B_i(t))^T$ is a vector-valued function of $B_i(t)$, which includes polynomial functions. Further, let $\Phi_{i,n} = \int_0^{t_n} \phi_i(B_i(s))ds$. Then,

$$\xi^g_{n,i} = \min_{\theta_i} \Phi^T_{i,n}(\theta_i - \theta_{i,0}) - \log \frac{\phi_i(B_i(\delta_i^{n_i}))^T\theta_i}{\phi_i(B_i(\delta_i^{n_i}))^T\theta_{i,0}}$$

which is a convex minimization problem and can be easily solved using gradient descent methods. Then, the G-CUSUM statistic h^g_n is then given as

$$h^g_n = \max\{0, h^g_{n-1} + \sum_{i=1}^{I}(\xi^g_{n,i} - \xi^g_{n-1,i})\} \qquad \text{where} \quad h^g_0 = 0.$$

The MLE-based likelihood ratio test can lead to poor change detection power when the dimensionality of Θ is large. To overcome this problem, a Lasso penalty term can be added for estimating the MLE of the service rate in each node, which reduces the dimensionality of changed parameters by inducing sparsity to the MLE in generalized likelihood ratio test [109, 83]. In large QNs, it is reasonable to assume that only a few service rates will deviate from the in-control values. The Lasso method induces sparsity in the estimated Θ, and therefore increases the probability to select the θ_i that change.

Adding a penalty term to MLE-based statistic is equivalent to maximize the penalized log-likelihood ratio, which is as follows

$$\begin{aligned}\xi^{\psi}_n =& \max_{\Theta} \left[\sum_{i=1}^{I}\sum_{n_i=1}^{D_i(t_n)} \log \frac{f_i(B_i(\delta_i^{n_i}), \Theta)}{f_i(B_i(\delta_i^{n_i}), \Theta_0)} - \int_0^{t_n} (f_i(B_i(s), \Theta) - f_i(B_i(s), \Theta_0))\, ds\right] \\ & - \psi||\Theta - \Theta_0||_1 \\ =& \sum_{i=1}^{I} \xi^{\psi}_{n,i},\end{aligned} \tag{3.4}$$

where I is the number of nodes in the QN, $i \in \{1, 2, \cdots, I\}$, and

$$\xi^{\psi}_{n,i} = \min_{\theta_i} \Phi^T_{i,n}(\theta_i - \theta_{i,0}) - \log \frac{\phi_i(B_i(\delta_i^{n_i}))^T\theta_i}{\phi_i(B_i(\delta_i^{n_i}))^T\theta_{i,0}} + \psi||\theta_i - \theta_{i,0}||_1.$$

Similarly, the test statistic for the P-CUSUM with a penalty ψ is defined as

$$h_n^{\psi} = \max\{0, h_{n-1}^{\psi} + \sum_{i=1}^{I}(\xi_{n,i}^{\psi} - \xi_{n-1,i}^{\psi})\} \qquad \text{where} \quad h_0^{\psi} = 0.$$

It is worth noting that the derivations of ξ_n, ξ_n^g, ξ_n^{ψ} in Equation (3.2), (3.3) and (3.4) show that our proposed log-likelihood ratio based CUSUM statistic only require departure timestamps and number of customers in for each queue in a QN. In the implementation of the CUSUM charts, the optimization step converges in a few iterations and did not pose numerical issues.

3.6 Numerical Study

In this section, the results of a simulation study to analyze the performance of the proposed CUSUM schemes in Section 3.5 are discussed. In the simulation experiments, a QN with ten nodes will be examined, which is shown in Figure 3.1. Each node of the QN consists of a single server. It is important to note that the likelihood-ratio-based CUSUM schemes are agnostic to the number of servers in the queue. Entities arrive at the first node and depart the system from the last node. All the servers are independent of each other and their service times are exponentially distributed with the rate $\mu_i, i \in \{1, 2, ..., 10\}$. The in-control values of service rate for each node are all set to be 1.1, that is $\Theta_0 = [1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1]$. The external arrivals to the first node are according to a Poisson process with rate $\lambda = 1$. This QN is equivalent to ten connected M/M/1 queues.

Monte Carlo simulations are used to analyze the ARL performance of our proposed CUSUM methods: SCUSUM, G-CUSUM and P-CUSUM charts. The designed out-of-control parameters of the service rates for SCUSUM chart is set as $\Theta_1 = 0.9\Theta_0 = 0.9 * [1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1, 1.1]$, which corresponds to a 10% decrease in service rates of all nodes. For the G-CUSUM, instead of a hypothesized change in service rate

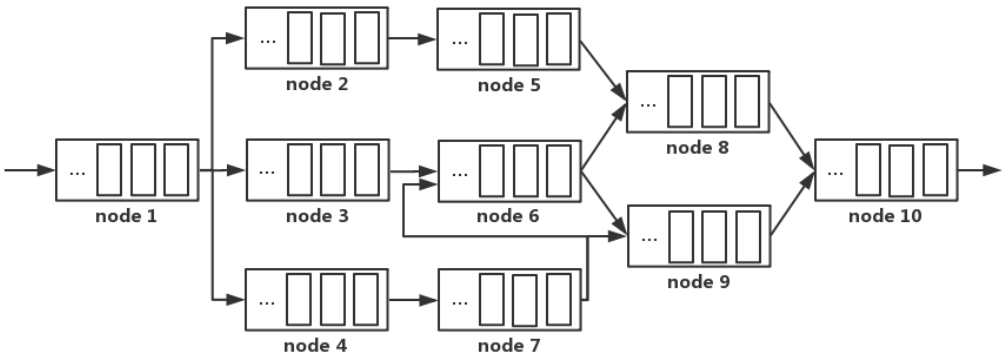


Figure 3.1: Structure of QN

for each node, the MLE of the service rate for each node should be computed to obtain the test statistics. The control limits for all methods are all set such that the $\mathrm{ARL}_0 = 100$.

We compare the performance of the proposed CUSUM scheme with two general multivariate SPC schemes: the multivariate CUSUM (MCUSUM) scheme [110] and the multivariate exponentially weighted moving average (MEWMA) chart [111] to monitor the queue length for every t_n, the time when n^{th} event occurs. This is consistent with previous literature on monitoring queue length of single server queues [55, 112]. Let $Q_n = [q_n^1, q_n^2, ..., q_n^{10}]^T$ denote the queue length for each of the ten nodes at t_n. The MEWMA and MCUSUM charts described here are meant to detect the change in service rate of the service nodes in the QN based on Q_n. The MEWMA test statistic for Q_n is defined as

$$T_n^2 = \frac{2-\gamma}{\gamma} Z_n^T \Sigma^{-1} Z_n$$

where $\gamma \in (0, 1]$ is a weighting parameter and Z_n is a vector calculated in a recursive form

$$Z_n = \gamma(Q_n - Q^0) + (1-\gamma)Z_{n-1}$$

where $Z_0 = 0$, and Q^0 and Σ are the mean and covariance of queue lengths under in-control scenario, which are estimated from 10,000 simulations of Q_n under the in-control setting. The recommended values of γ is between 0.05 and 0.2 [113]. In the reported results $\gamma = 0.2$ was found to be the best for detecting small changes.

The MCUSUM statistic is defined as:

$$MC_n = \max\{0, \sqrt{D_n^T \Sigma^{-1} D_n} - \tilde{k}\omega_n\},$$

where

$$D_n = \sum_{i=n-\omega_n+1}^{n} (Q_i - Q^0)$$

and

$$\omega_n = \begin{cases} \omega_{n-1} + 1 & \text{if} \quad MC_{n-1} > 0 \\ 1 & \text{otherwise} \end{cases}.$$

Also, following the recommendations in [110] and the $\Theta_1 = 0.9\Theta_0$ values, $\tilde{k} = 0.12$ was selected.

3.6.1 ARL Comparisons for Detecting the Change of All Service Nodes

Figure 3.2 presents the ARL comparisons for detecting the decrease in service rates of all nodes ranging from Θ_0 to $0.4\Theta_0$. The comparison shows that our proposed CUSUM charts significantly outperform the MEWMA and MCUSUM charts. Among them, the G-CUSUM is comparable to the SCUSUM in terms of the ARL performance and exhibits better sensitivity than the P-CUSUM chart in detecting the small changes of all service nodes. The Lasso penalty would force some of estimated service rates equal to their in-control values, which is not consistent with the fact that all service rates have been changed. Hence, P-CUSUM chart is less effective when all the nodes of the QN have slowed down.

On the other hand, when the actual service rates are much slower than the designed out-of-control parameter Θ_1, the performance of G-CUSUM chart deteriorates and SCUSUM

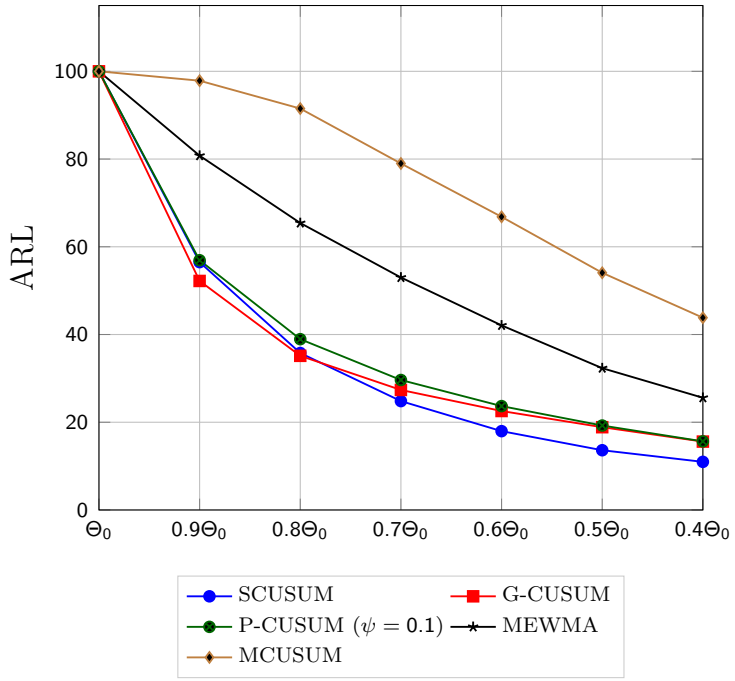


Figure 3.2: ARL comparisons in detecting the decrease of the service rates of all nodes

chart is still sensitive. It is due to the fact that simple CUSUM charts are usually effective when actual change direction is similar to the hypothesized out-of-control change [104]. On the other hand, when service rates decrease, the number of departure events decreases and the estimation error in G-CUSUM test statistic increases, which can explain the slight decrease in performance for smaller values of Θ observed in Figure 3.2.

3.6.2 ARL Comparisons for Detecting the Change of Single Node

For detecting change of the single node in the QN, ARL comparisons are discussed for the first node, middle node and last node. Figure 3.3 and Figure 3.4 show the ARL comparisons in detecting the decrease of the service rate of the first node and last node, respectively. Firstly, we demonstrate that the proposed CUSUM charts perform much better when compared to MEWMA and MCUSUM charts. Because the change in queue length of first node or last node significantly dominates that of other nodes when we only decrease the service rate of the first or last node, it makes MEWMA and MCUSUM less sensitive to the slowing in the service rate of other nodes. Among all the CUSUM charts, the G-CUSUM

and P-CUSUM charts are more sensitive than SCUSUM chart in detecting any amount of decrease of the service rate for the first node and last node. However, this is not surprising. The designed out-of-control parameters of SCUSUM charts assume all the service rates have reduced, therefore its relatively poor ability to detect the change of service rate of a single node in comparison with G-CUSUM and P-CUSUM charts. Furthermore, it reveals that the G-CUSUM chart is more sensitive than the P-CUSUM chart. But the first and last node of the network are different than the other nodes. Change in service rate of either changes the performance of the whole QN. So, the need for detecting a sparse change, which is the goal in P-CUSUM chart, is not realized. Indeed the bias resulting from penalizing the likelihood could also impact the performance of P-CUSUM chart.

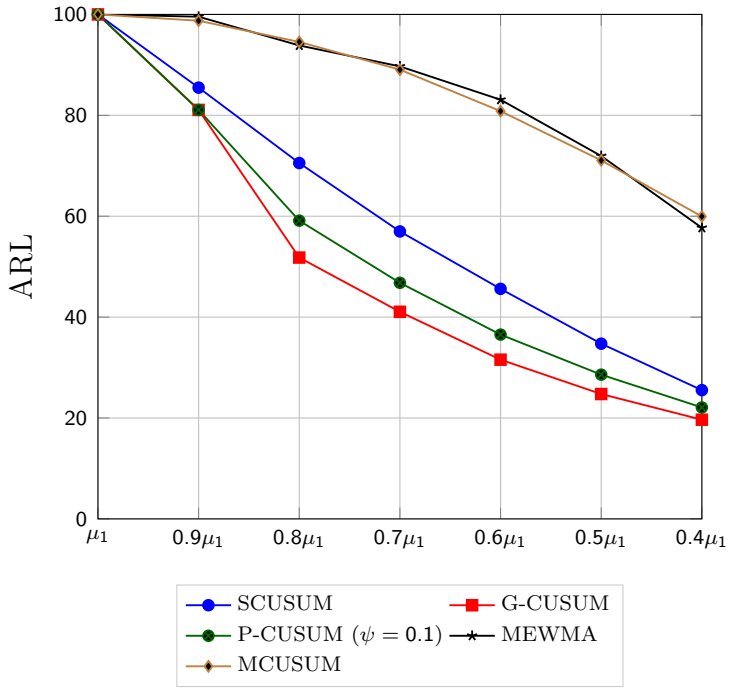


Figure 3.3: ARL comparisons in detecting the decrease of μ_1

Figure 3.5 shows ARL comparisons for various monitoring schemes in detecting the decrease of the service rate of the fifth node, which is located in the middle of the network. Again, Figure 3.5 shows that MEWMA and MCUSUM perform poorly. Also, G-CUSUM and P-CUSUM charts have better performance in detecting the change in service rate of a

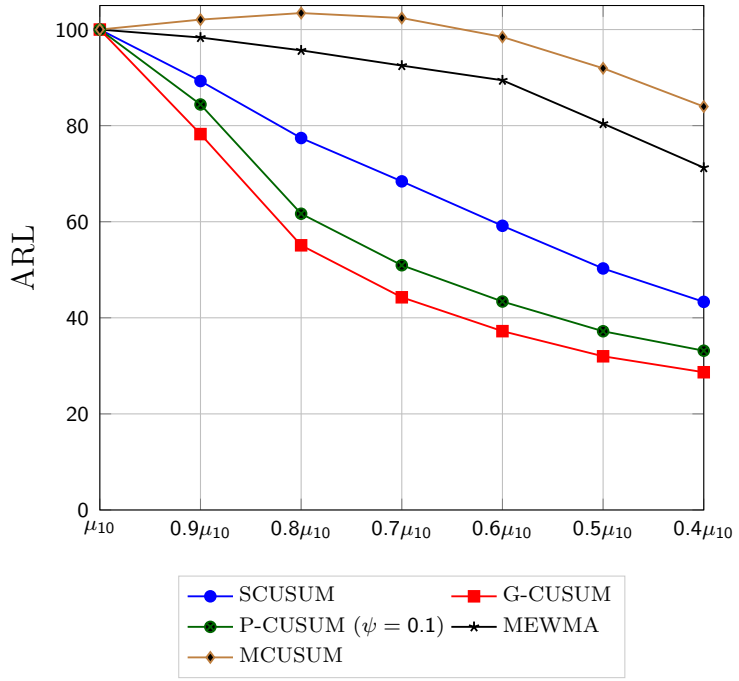


Figure 3.4: ARL comparisons in detecting the decrease of μ_{10}

single node than SCUSUM. In addition, it is observed that the P-CUSUM chart is more effective in detecting a small decrease of the service rate in fifth node and the G-CUSUM chart exhibits better sensitivity in detecting the moderate and large decreases in service rate of a single node. The latter observation is consistent with the findings in Figure 3.3 and Figure 3.4. Therefore, if the objective is to detect a small change in the service rate of a single node using a small sized sample, adding a penalty term is recommended.

3.6.3 Identity the Exact Out-of-control Node Using Penalized CUSUM Chart

Traditional multivariate SPC scheme like MEWMA and MCUSUM control chart statistics are computed based on the covariance matrix in data, they can be used to detect the potential change for multivariate process but not to identify which variate has changed, but the latter is more important for quality control practitioners. Thus, our proposed penalized CUSUM charts can overcome the limitation to identify the exact out of control node when only single node in a queueing network has changed. The designed penalized CUSUM charts can return us a set of estimated departure rates for each node that has potentially changed,

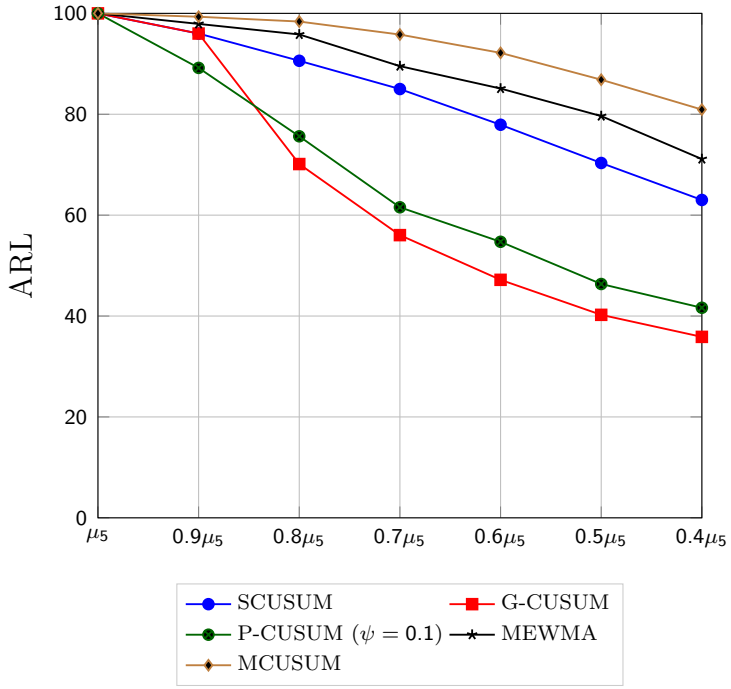


Figure 3.5: ARL comparisons in detecting the decrease of μ_5

then the node with the minimal estimated departure rate is signaled as out of control. In order to verify the efficiency of penalized CUSUM chart, a parallel network with 5 nodes is examined, which is shown in Figure 3.6.

Table 3.1 is the accuracy of P-CUSUM for identifying node 2 as out-of-control if only node 2 has decreased. It shows that the accuracy is increasing when the degree of change for node 2 is increasing. The accuracy is defined as the probability of identifying node 2 as out-of-control within all the five nodes during a process.

Table 3.1: The accuracy of P-CUSUM for identifying node 2 as out-of-control if only node 2 has decreased

Actual change of μ_2	Accuracy of P-CUSUM
$0.9\mu_2$	30%
$0.8\mu_2$	40%
$0.7\mu_2$	52%
$0.6\mu_2$	65%
$0.5\mu_2$	76%
$0.4\mu_2$	85%

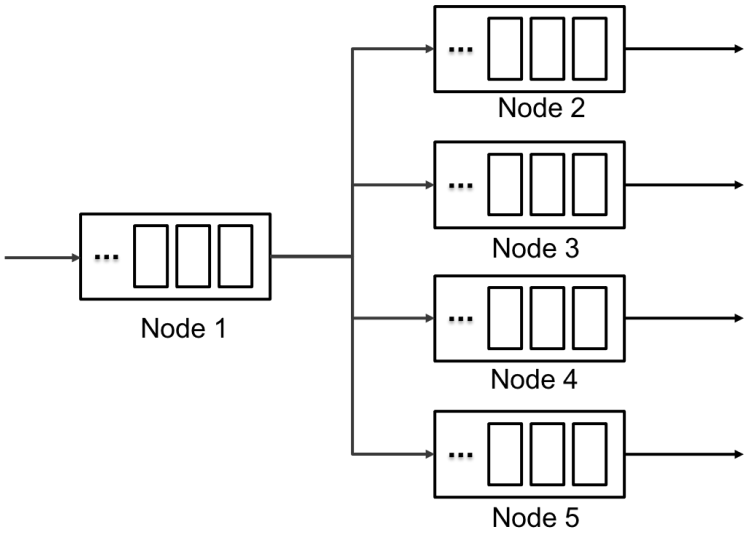


Figure 3.6: Parallel queueing network

3.7 Case Study: Monitoring the Flow of Patients in an ED

In this section, the proposed monitoring schemes are evaluated to monitor the flow of patients in the emergency department of a major academic medical center. The patient flow in the ED is modeled as a QN with five nodes, which are registration desk and four clusters of beds with a team of healthcare providers in each cluster. They are called East, Center 1, Center 2 and West nodes. These correspond to the distinct pods in the ED from where the data is collected and illustrated in Figure 3.7. Patients visiting the ED wait until they are assigned to different wards. The patients leave the ED (admitted to the hospital wards or discharged) after being served in the wards. We are interested in monitoring the service rate of these five nodes, shown in Figure 3.7.

Here we assume each service node in ED is a processor-sharing queue with a state-dependent service rate function, because many existing papers show that a processor sharing queue with state-dependent service rate function has more flexibility to model complicated healthcare systems like ED while approximating the actual system performance reasonably well [96]. Figure 3.8 illustrates an example for the empirical distribution of the patient occupancy levels in one of the nodes at our partner ED, e.g. the east node. The occupancy

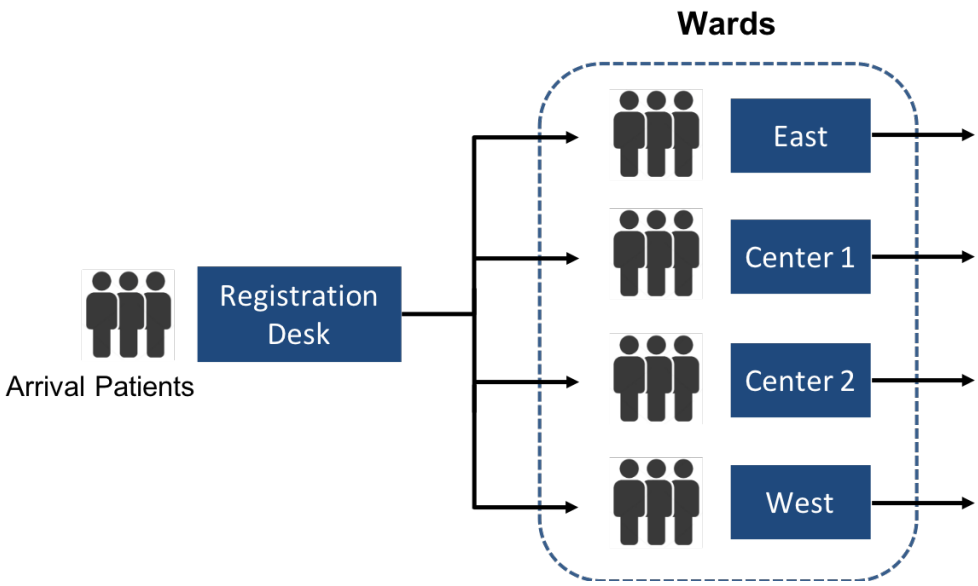


Figure 3.7: Patient visit flow of the emergency department (ED) of a large academic medical center

level at a given time represents the total number of patients in the node. In the figure, The x-axis is the state namely the number of patients in the east node, and the y-axis corresponds to the frequency of the state. We find that assuming a processor-sharing queue with a state-dependent service rate for our partner ED can best replicate the empirical occupancy distribution curve compared to the conventional M/M/1 queue, which clearly deviates from the empirical distribution. We believe that the processor-sharing queue provides a better fit, since the ED is a complex service environment with many shared resources (nurses, doctors, medical equipment, labs, etc) and multitasking situations. For example, it is natural to postpone the treatment of a low-risk patient for a newly arriving high-risk patient. Also, a doctor or nurse can be acted as a single server to deal with multiple patients simultaneously. Thus, processor-sharing queue is more appropriate to fit these complex situations encountered in the ED service environment, while traditional queuing models, in which the server focuses on servicing just one patient at a time, are not flexible enough to deal with these complexities. Because both resource sharing and multitasking mechanisms that commonly

seen in the ED service environment are basically a processor sharing framework in which all patients currently in the ED gain equal attention from each medical staff [114].

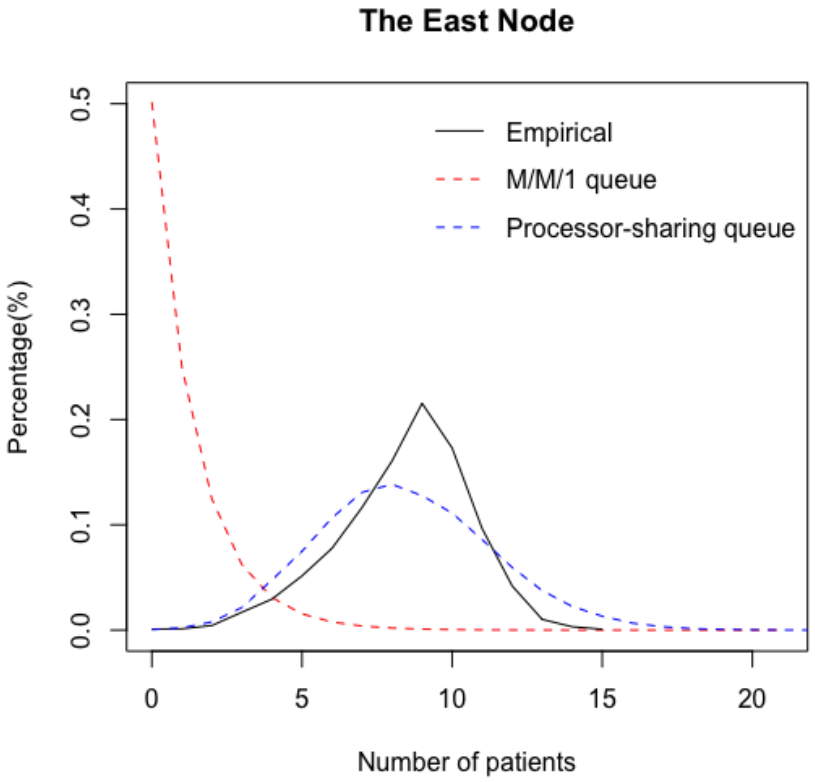


Figure 3.8: Histogram of patient occupancy of the east node in ED

We select the first 183 days in Year 2016 as training data to estimate the in-control departure rate using model fitting methods. We adopted a linear form with $\mu_i(t) = \theta_i f_i(n_i(t))$ to define the service rate of node i in ED, where $n_i(t)$ denotes the number of patients in node i at time t and θ_i is a parameter corresponding to f_i, and f_i represents a transform of $n_i(t)$ such as the logarithm, square root, square or cube of $n_i(t)$. Then, 10-fold cross validation (CV) method is applied on the in control data to select the best model. Table 2 is the CV errors for different models. Among different models, we can see that Model 3 (a linear function with respect to the square root of number of patients) produces the minimum errors, which is then chosen as the best model to explain the relationship between

total service rate and the number of patients for each node. As a result,

$$\mu_1(t) = 203.1\sqrt{n_1(t)}, \mu_2(t) = 49.93\sqrt{n_2(t)}, \mu_3(t) = 47.43\sqrt{n_3(t)},$$
$$\mu_4(t) = 51.2\sqrt{n_4(t)}, \mu_5(t) = 51.61\sqrt{n_5(t)},$$

Table 3.2: CV errors for different models

	Model 1 $\mu_i(t) = \theta_i n_i(t)$	Model 2 $\mu_i(t) = \theta_i \log(n_i(t))$	Model 3 $\mu_i(t) = \theta_i \sqrt{n_i(t)}$	Model 4 $\mu_i(t) = \theta_i n_i^2(t)$	Model 5 $\mu_i(t) = \theta_i n_i^3(t)$
Node 1	755.9	746.7	**702.6**	796.6	811.8
Node 2	299.2	299.2	**299.2**	301.1	304.4
Node 3	293.2	293.2	**293.2**	295.1	298.6
Node 4	290.6	289.9	**289.7**	294.6	298.8
Node 5	336.3	336.2	**336.2**	339.6	345.4

We use the last 183 days in Year 2016 as the test data set. The control limits are set as the 90% percentile of the test statistics for the training data. The proposed CUSUM charts are used to detect the decrease in the service rates and then the MEWMA and MCUSUM charts are used to compare with proposed CUSUM charts. The number of days classified as in control and out of control are presented in the confusion matrices in Table 3-6. Table 3 shows that there are total 51 days in the test dataset are labeled as out of control by SCUSUM chart, while 36 days among the testing set are signed as out of control by MEWMA chart, in which 21 days are identified as out of control by both SCUSUM and MEWMA charts, and 30 days are identified as out of control by SCUSUM chart only. Similar results for the comparisons for P-CUSUM and MEWMA charts are given in Table 4, SCUSUM and MCUSUM charts are given in Table 5, and P-CUSUM and MCUSUM charts are given in Table 6.

Table 3.3: Confusion matrix for SCUSUM and MEWMA charts

		SCUSUM		
	$N = 183$	Out of control	In control	Total
MEWMA	Out of control	21	15	36
	In control	30	117	147
	Total	51	132	

Table 3.4: Confusion matrix for P-CUSUM and MEWMA charts

		P-CUSUM		
	$N = 183$	Out of control	In control	Total
MEWMA	Out of control	20	16	36
	In-control	26	121	147
	Total	46	137	

Table 3.5: Confusion matrix for SCUSUM and MCUSUM charts

		SCUSUM		
	$N = 183$	Out of control	In control	Total
MCUSUM	Out of control	33	26	59
	In control	18	106	124
	Total	51	132	

Table 3.6: Confusion matrix for P-CUSUM and MCUSUM charts

		P-CUSUM		
	$N = 183$	Out of control	In control	Total
MCUSUM	Out of control	29	30	59
	In control	17	107	124
	Total	46	137	

In the test dataset, there are 15 days signaled out-of-control by both SCUSUM and P-CUSUM but not the MEWMA and MCUSUM charts. To get further insight into the reason for this, we study October 20, 2016 in further detail. Our methods found there is an overall decrease in service rate for all the nodes on Oct 20. However, the east node has decreased dramatically compared to other nodes. Then we compare the actual departure rate with the in-control departure rate for the east node on Oct 20 in Figure 3.9. This figure shows the departure rate has clearly dropped at every time of the day for the east node on Oct 20, 2016.

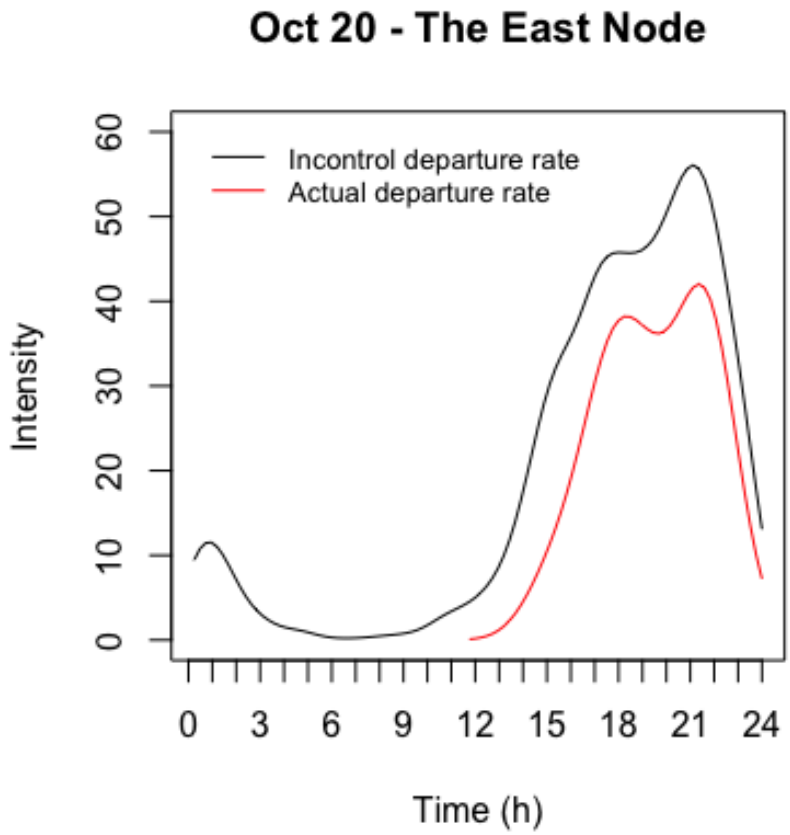


Figure 3.9: East node departure intensity comparison on Oct 20, 2016, which was signaled out-of-control by both SCUSUM and P-CUSUM but not the MEWMA and MCUSUM

Table 3.7 shows actual average queue length, and their in-control values for each node on October 20, 2016. We can observe that, except for the registration node, the average

queue length of all the other nodes on October 20 just slightly deviated from the in-control average value, hence, the MEWMA and MCUSUM test statistics were not able to distinguish them as out of control. However, as shown in Figure 3.9, the service rate for the east node had clearly decreased on October 20, 2016 and we are able to detect it using the proposed methods.

Table 3.7: The average queue length comparisons on October 20, 2016

		Registration	East	Center 1	Center 2	West
Ave. Queue Length	October 20	7.8	3.17	8.32	9.55	9.64
	In-control	4.31	4.25	8.67	8.02	9.68

The analysis of the real data leads to an important conclusion that monitoring the service rate in the ED is needed. It also shows that traditional performance measures of queuing system such as queue length often are unable to reflect the service ability in ED. The result can assist operations managers to improve the timeliness of care in the ED. The proposed methods can be used as a retrospective evaluation tool. If a specific day for a node is signaled as out of control, the operations managers in the ED would retrospectively look into probable causes of the alarm and take necessary action to resolve it. For example, if we found a day of the week is signaled as out of control frequently, redesigning the weekly staffing schedule on that day could be considered.

3.8 Conclusion

In this paper, we propose new CUSUM control charts based on count data to monitor the service rates of a QN with state-dependent queues. The proposed CUSUM charts are compared with the MEWMA and MCUSUM charts using the ARL criteria to detect out-of-control scenarios. A major contribution of this research is the development of an easy to implement and efficient likelihood-ratio-based CUSUM charts, G-CUSUM and P-CUSUM charts for monitoring QNs, which could overcome the limitation of the normality assumption and do not need know the potential change in service rate of the queueing nodes in a QN,

and thus have important practical applications. Numerical studies based on a simulated QN demonstrated that the proposed CUSUM charts can outperform traditional approaches on a variety of out-of-control scenario detection tests. Further, a case study focusing on monitoring the daily patient flow of an ED demonstrates the efficacy of the proposed methods in a real application.

There are several extensions of the methods developed in this paper. The current monitoring scheme is based on the likelihood ratio statistic, which requires the sample path can be observed completely. However, there are some challenges associated with obtaining the likelihood ratio statistic when only limited and partial samples can be observed. Generalizations and extensions of this method to study problems such as changes in optimal routing policies and dependence on factors external to an ED that can cause delays in the ED are part of our ongoing research. Another important extension of this paper could involve the study of approximation methods in establishing theoretical understanding of statistical monitoring of QNs. Specifically, the application of diffusion approximation methods can help establish theoretical performance guarantees of CUSUM methods developed here. In addition, other than the real case application in ED, monitoring the patient flows in other units, such as the intensive care units is important for further methodological development and application of the proposed methods in quality control of healthcare systems.

Chapter 4: Statistical Monitoring the Cascade of Care for Patients with Alcohol Use Disorder [1]

4.1 Overview

Recently, a Cascade of Care (COC) framework has been widely applied to improve system-level practice and treatment outcomes for various chronic medical conditions. Increasingly, the CDC and NIDA are suggesting this framework as it can be used to trace and evaluate the treatment progress and outcomes at both the individual patient level and population level. However, very limited research has been conducted on COC development for alcohol use disorder (AUD). This paper aims to develop and test a model for measuring and monitoring the treatment processes of AUD using a COC framework. First, an innovative continuous-time stochastic process model is proposed to represent the dynamics of the COC for AUD treatment, from which benchmarks for COC can be developed by learning ideal patterns during different stages in care for AUD related to outcomes that indicate improved health. To the best of our knowledge, this study would be the first extension of the continuous-time stochastic modeling approach to AUD treatment processes. Then, a new statistical monitoring scheme is developed to identify the patients whose care deviated from the baseline model. The efficacy of the proposed method is demonstrated by simulations and a real case study focusing on monitoring the patient's follow-up visit after initiating the treatment for AUD. Finally, the key factor affecting treatment outcome is identified,

[1] Portions of Chapter 4 have been reproduced by permission for dissertation use only, "A Continuous-time Stochastic Modeling Approach for Monitoring the Cascade of Care for Patients with Alcohol Use Disorder" in Proceedings of the 2022 IISE Annual Conference, Institute of Industrial and Systems Engineers [115].

which would help clinicians or public health associations develop subsequent interventions to improve treatment outcomes.

4.2 Introduction

Alcohol use disorder (AUD) is a chronic disease characterized by compulsive or uncontrollable alcohol use despite harmful consequences and long-lasting changes in the brain. According to the 2019 National Survey on Drug Use and Health, 14.5 million people ages 12 and older had AUD. This number includes 14.1 million adults and 0.4 million adolescents. The rate of all alcohol-related hospitalization increased 47 percent between 2006 and 2014 [116] and alcohol becomes the third leading preventable cause of mortality in USA, causing more than 95,000 alcohol-related deaths annually [117]. Also, AUD costs more than $249.0 billion annually which results in a big economic burden in the United States [118]. Since AUD is very harmful to individual's health and the society, effective treatment for AUD is needed. Due to the long-lasting changes in the brain caused by AUD, relapse is common, so continuity of care is the critical factor for successful treatment for AUD. However, the majority of the patients who are identified with AUD do not initiate treatment [68]. And for those who initiated treatment, fewer than 15% continued in treatment. Therefore, it is important to measure and monitor the treatment process for AUD to identify care processes that lead to successful outcomes and patients whose linkage to care failed to occur in a timely manner and led to negative outcomes.

Although monitoring the treatment process for AUD has not been studied, some papers have developed approaches for monitoring and improving the treatment outcome for other substance use disorder such as opioid use disorder (OUD). For example, Matteliano et al [72] developed a biopsychosocial-spiritual assessment model which is a comprehensive approach for monitoring and improving the adherence treatment of chronic opioid therapy for patients with persistent pain. Manchikanti et al [73] proposed a evaluation tool including a chart review to monitor controlled substance intake for patients with chronic pain, which results

in 50% reduction in opioid abuse. These researches have successfully improved the access to care for OUD patients, but not continuity of care or outcomes despite an increase in public and private expenditures. In the meantime, although access may have improved for the treatment of OUD, it has not increased for people with other substance use disorders. Thus new approach is needed to monitor the quality of system-level care in substance use disorder treatment especially for the alcohol use disorder (AUD) that is currently lack of attention.

Cascade of Care (COC) is a whole-system approach to assess the effectiveness of treatment process for various health conditions. Increasingly, the CDC and NIDA are suggesting this framework as it can be used to trace and evaluate the treatment progress and outcomes at both the individual patient level and population level [8]. Treatment cascades measure patient flow through the system and can be used to identify process breakdowns missed by single care stage like initiation or engagement that are required to achieve a successful treatment outcome [119]. The COC model has been widely used in assessing the effectiveness of treatment systems in HIV, HCV, diabetes, and other conditions [9]-[11], but very limited work has been published on COC development for AUD. To fill this gap, this paper aims to develop and test a model for measuring and monitoring the treatment processes of AUD under a COC framework. First, an innovative continuous-time stochastic process model (called CTCOC model) is proposed to represent the dynamic of the COC for AUD treatment, from which benchmarks for COC can be developed by learning ideal patterns during different stages in care for AUD related to outcomes that indicate improved health. To the best of our knowledge, this study would be the first extension of the continuous-time stochastic modeling approach to AUD treatment processes. Furthermore, a new statistical monitoring scheme is developed to effectively identify the patients whose care deviated from the baseline based on the CTCOC model.

The remainder of this paper is organized as follows: The general formulation for CTCOC model is provided in Section 4.3. Section 4.4 introduces the statistical monitoring scheme that we developed to monitor the patient COC events. Section 4.5 presents a simulation

study to demonstrate the efficiency of the proposed method by comparing it with the traditional method. Section 4.6 presents the real case study focusing on monitoring the follow up visit for AUD patients after initiating treatment. Section 4.7 identifies the key factor affecting the patient outcome, and conclusions are conducted in the last section.

4.3 Continuous-time COC (CTCOC) Model

Consider a COC consisting of a set of events, $E_0, E_1, E_2, \ldots, E_n$, where n is the total number of unique events that can be obtained for an AUD treatment process. For example, E_0 can be considered as the identification of AUD, E_1 can be the initiation of treatment, E_2 can be the first follow-up appointment with a physician or therapist, and so on. Data for individual patient undergoing AUD treatment can be represented as:

$$(E_0^i, t_0^i), (E_1^i, t_1^i), \ldots, (E_{n-1}^i, t_{n-1}^i), (E_n^i, t_n^i)$$

where the index i denotes the ith patient, $t_0^i < t_1^i < \cdots < t_n^i$ and $E_0^i, E_1^i, \ldots, E_n^i$ are successive COC events-logs data for patient i. For every pair of E_p and E_q, if E_q immediately follows E_p for patient i, the time of occurrence of E_q is denoted as $s_{p,q}^{i,j}$, where j denotes that the jth time E_q immediately follows E_p. Specifically, if $E_{j-1}^i = E_p$ and $E_j^i = E_q$, then $s_{p,q}^i = t_j^i$. This is illustrated in Figure 4.1

For every patient i and every pair of events E_p and E_q, a counting process can be defined as follows

$$N_{p,q}^i(t) = number \quad of \quad s_{p,q}^{i,j} \leq t.$$

$N_{p,q}^i(t)$ indicates the number of times the event E_q is recorded immediately after E_q for patient i by the time t. The CTCOC model is based on the assumption that the rate of occurrence of E_q after E_p in the ideal subpopulation specific to an AUD satisfies the following:

$$Prob(N_{p,q}^i(t + dt) - N_{p,q}^i(t) = 1) = r_{p,q}(t)dt, \tag{4.1}$$

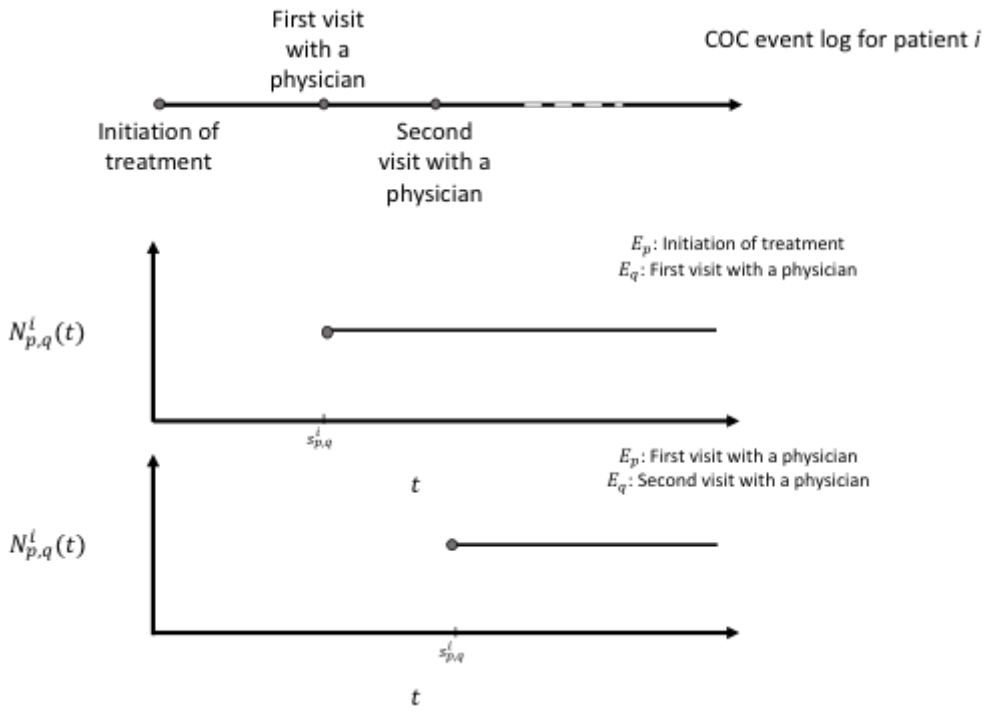


Figure 4.1: CTCOC model for COC event log

where Prob(·) denotes the probability of the event within the parenthesis. In stochastic process literature, $r_{p,q}(t)$ is referred to as the intensity of occurrence of E_q following E_p , and

$$R_{p,q}(t) = \int_0^t r_{p,q}(s)ds$$

is the cumulative intensity of occurrence of E_q following E_p. For the counting process $N^i_{p,q}(t)$ that satisfy the assumption mentioned in equation 4.1, it is known that the expected value of $N^i_{p,q}(t)$ is $R_{p,q}(t)$. This concept is illustrated from an example in Figure 4.2. If I denotes the set of all such patients in the ideal subpopulation, $R_{p,q}(t)$ can be estimated as

$$\hat{R}_{p,q}(t) = \frac{1}{N_I} \sum_{i \in I} N^i_{p,q}(t),$$

where $\hat{R}_{p,q}(t)$ is the estimated value of $R_{p,q}(t)$ and N_I the total number of patients in the set I. This approach is a nonparametric method of estimating the cumulative intensity function

$R_{p,q}(t)$. To the best of our knowledge, the proposed project would be the first extension of the continuous-time stochastic modeling approach to AUD treatment processes.

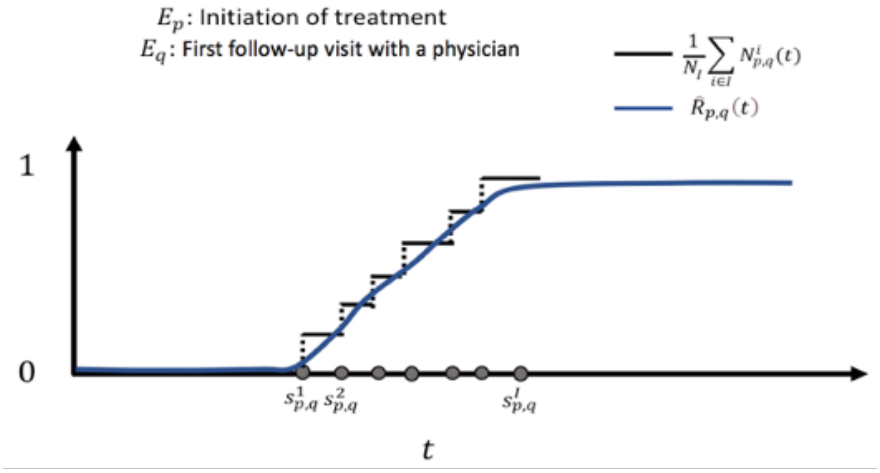


Figure 4.2: Rate of occurrence of E_q after E_p

4.4 Statistical Monitoring Scheme

This section develops a statistical monitoring scheme that detects changes in the transition rate between different stages along COC, in real-time, whether a patient adheres to the ideal treatment process or not. A monitoring scheme often involves two steps. The first step is to find the in control transition intensity between the COC events, the second step is to signal a patient as out-of-control if his or her care deviates from the in-control transition intensity. The in-control or ideal patients could be based on specific outcomes y_i. For example, if y_i denotes the number of hospitalizations in a year, $y_i < 1$ could be used to select the in-control patient COC event group. Let $R^0_{p,q}(t)$ be the in-control cumulative intensity of occurrence of E_q following E_p, which is estimated based on the in control patients, and $R^1_{p,q}(t)$ be the actual intensity of a test patient. If $R^0_{p,q}(t) \neq R^1_{p,q}(t)$, the test patient is said to be "out-of-control". The purpose of the statistical monitoring scheme described here is to detect such a change. For each patient and time point t, a test statistic d(t) is defined as to test the following hypothesis

$$H_0 : R^0_{p,q}(t) = R^1_{p,q}(t)$$

$$v.s.$$

$$H_1 : R^0_{p,q}(t) \neq R^1_{p,q}(t)$$

Then a decision rule is defined as

$$d(t) \geq h(t)$$

such that

$$P(d(t) \geq h(t)|H_0) = \alpha,$$

where h(t) is the "out-of-control" signal threshold and α is a user specified type I error rate, which is typically set as 10%.

The proposed CTCOC model allows us to develop a metric for measuring how much an individual patients' care deviates from ideal. As an example, consider the case where E_p is the initial treatment of an AUD and E_q is the subsequent visit with a healthcare provider. If a patient delays his or her first follow-up visit to healthcare provider, $s^i_{p,q}$ for this patient would be larger. Thus, the area between the curves $R_{p,q}(t)$ and $N^i_{p,q}(t)$ will be large. This is illustrated in Figure 4.3. Hence, we can define a test statistic $d^i_{p,q}(t)$ as follows:

$$d^i_{p,q}(t) = \int_0^t |\hat{R}_{p,q}(s) - N^i_{p,q}(s)|ds.$$

The average value of $d^i_{p,q}(t)$ for a subgroup of patients can provide a measure of deviation of the subgroup from the ideal $R_{p,q}(t)$. Similarly, the mean square of $d^i_{p,q}(t)$ for various COC events E_p and E_q can give combined metric for multiple pairs of COC events. $d^i_{p,q}(t)$ is similar to the Kolmogorov-Smirnov test statistic, which can be used to develop statistical tests to measure the statistical significance of the deviation from the ideal COC. The proposed the statistical test is then compared with the traditional goodness-of-fit test method for

continuous distributions, e.g. Chi-squared test, where the test statistic is defined as

$$S^i_{p,q}(t) = \sum_{k=1}^{K} (O^{i,k}_{p,q} - M^k_{p,q})^2/(M^k_{p,q}),$$

where K indicates the total number of bins that the continuous distribution has been discretized, $O^{i,k}_{p,q}$ is the observed number of transitions from E_p to E_q for patient i at each segment, and $M^k_{p,q}$ the expected number of transitions from E_p to E_q for each segment.

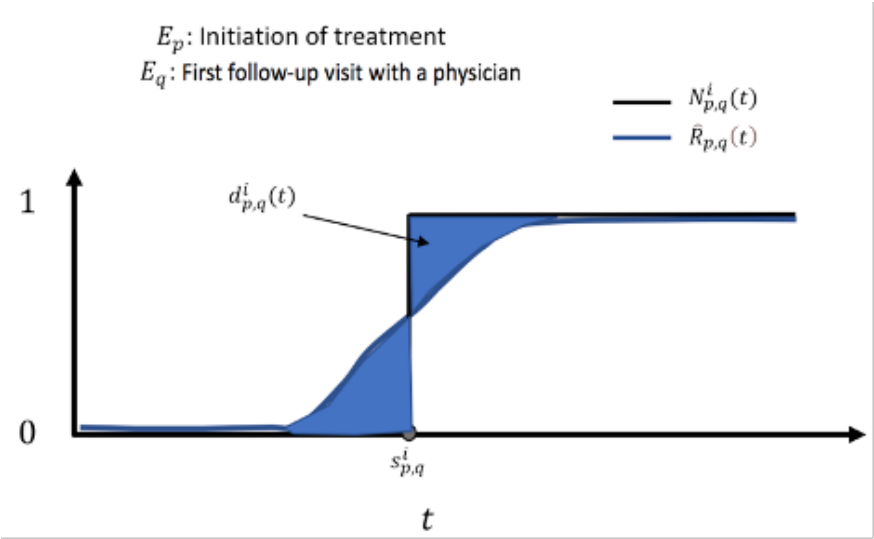


Figure 4.3: Measure of deviation from ideal COC

4.5 Simulation Study

In this section, a simulation study focuses on a two-state CTMC model is presented to demonstrate the efficacy of the proposed statistical monitoring scheme. Considering a two-state CTMC with states indexed as 1 and 2. The initial state at time t=0 can be either 1 or 2 with an equal probability. The transition rate matrix for the CTMC for $0 < t < 1$ is given as

$$Q(t) = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix},$$

where $\lambda_{11} = -\lambda_{12}$ and $\lambda_{22} = -\lambda_{21}$. Here we let $\lambda_{12} = \theta - 10\sin(2\pi t)$ and $\lambda_{21} = \theta$. In the simulation study reported here, we are interested in detecting changes in λ_{12}, where the transition rate λ_{12} is time-inhomogeneous and always greater than or equal to zero. Here we set $\theta = 20$ when λ_{12} is under control, since $\theta = 20$ can make sure λ_{12} is always a positive value. A change in θ represents an out-of-control scenario of the transition rate λ_{12}. The larger difference between the actual and in-control value of θ means the out-of-control scenario for λ_{12} is more deviated from the in-control scenario.

In this paper, we use Type II error rate, also known as misdetection rate if the system is known as out-of-control, to examine the performance of our proposed control chart compared with a traditional Chi-square chart. It is defined as the probability of failing to give an out-of-control signal when a system is actually out of control. First, we specify the type I error rate α as 10%, which can transfer to a type II error rate with 90% when θ is in-control. We use this value to find the control limit and we expect the type II error rate for detecting a system that is actually out of control is lower this value. The smaller the value of Type II error rate for a particular change, the greater the efficiency of the chart to detect the change. Figure 4.4 and Figure 4.5 show the Type II error rates for detecting changes in θ for our proposed control chart and the traditional Chi-square chart. We can see that the proposed control chart performs better than Chi-square chart for detecting any direction of change in θ or λ_{12}. In particular, for detecting the decrease of θ (which is equivalent to detecting the delay of the transition from state 1 to state 2), the figures show that our method is much more sensitive in detecting the delay of the transition compared to the Chi-square chart, while the latter can hardly detect any delay especially for smaller magnitude of decrease in θ.

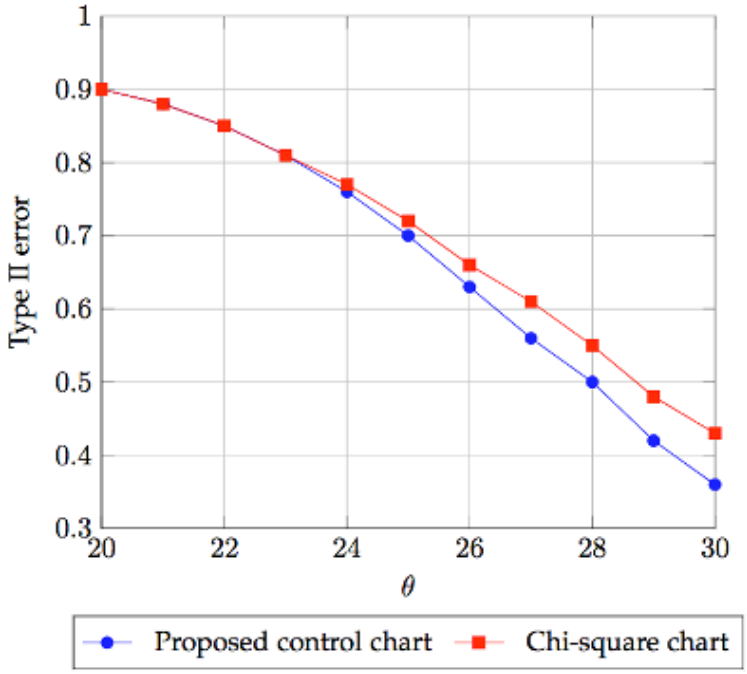


Figure 4.4: The type II error rates for detecting increase in θ

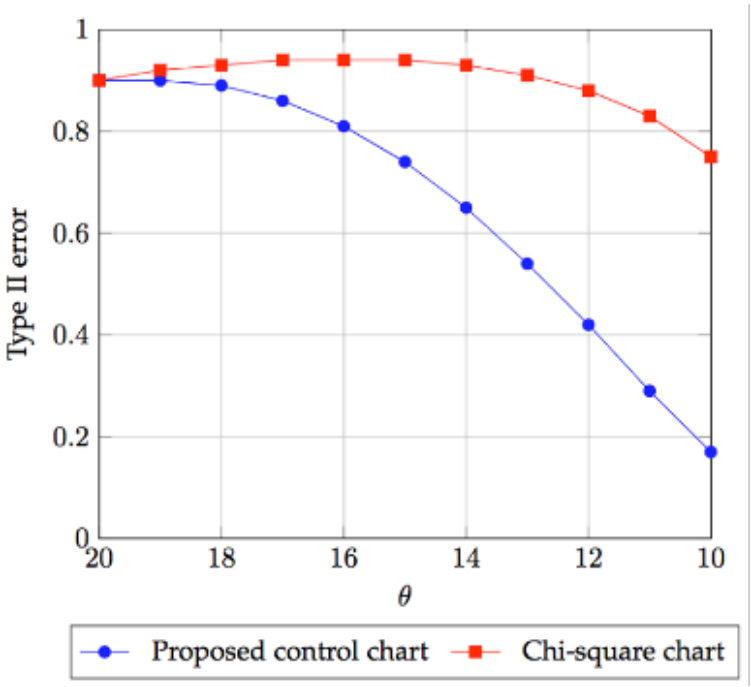


Figure 4.5: The type II error rates for detecting decrease in θ

4.6 Real Case Study

4.6.1 Data Description

In this section, the proposed monitoring scheme is evaluated on the DE-SynPUF data obtained from the website of the Centers for Medicare and Medicaid Services. This particular dataset provides a set of synthetic claims data from 2008 to 2010, including inpatient, outpatient and carrier claims. The claims ICD-9 diagnosis codes (e.g. 303.9x, 303.0x, 305.0x, etc.) are used to identify whether the visit is related to AUD treatment, and the patients with greater than 2 visits related to AUD treatment are adopted for analysis. First, we need to identify in-control or ideal patients based on favorable treatment outcomes. These treatment outcomes could be discrete-values, such as number of ED visits or number of hospitalizations, or continuous-valued such as cost of care. Here we consider the AUD patients with no hospitalizations record during treatment as in control samples and the rest are treated as testing data. Based on the way the training and testing data are split, there are 3397 in-control training samples and 2736 testing samples used in the case study.

We assume all these patients are identified or diagnosed as AUD on 2008/01/01, and the first claim record related to AUD treatment is considered as the initiation visit of the treatment. We are interested in monitoring the patients' follow-up visit with a doctor after initiating the treatment for AUD. Figs. 4 (a) and (b) show boxplot of the days spent and the cumulative intensity of occurrence for the follow-up visit after initiating the treatment for all in-control data, respectively. We can see the first quantile and the third quantile are 52 and 275 days, respectively. Also, the in-control median and average time spent for a patient to seek for a follow-up doctor's visit is 143 and 182 days, respectively. It may be due to the fact that the physicians often recommend 180-day duration on average for receiving medications, which can be self-managed at home, after the initial doctor's visit to achieve beneficial treatment outcome.

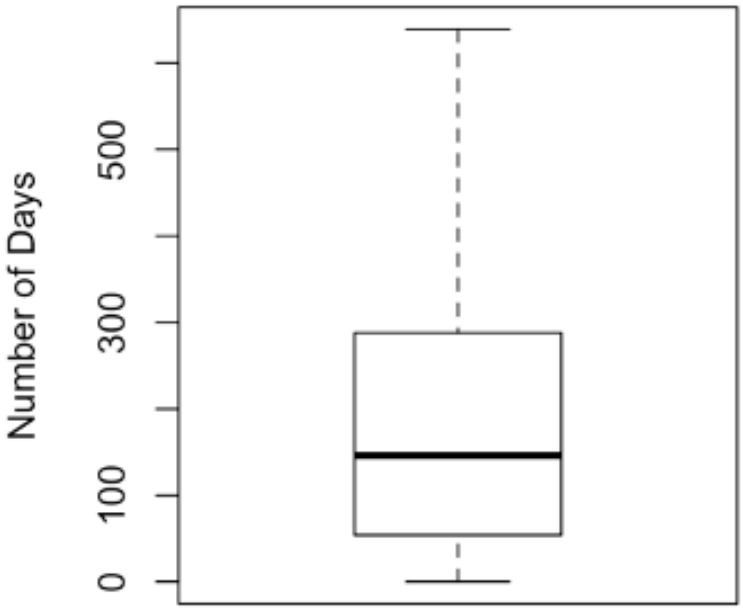


Figure 4.6: The boxplot of the days spent for the follow-up visit after initiating the treatment for all in-control data

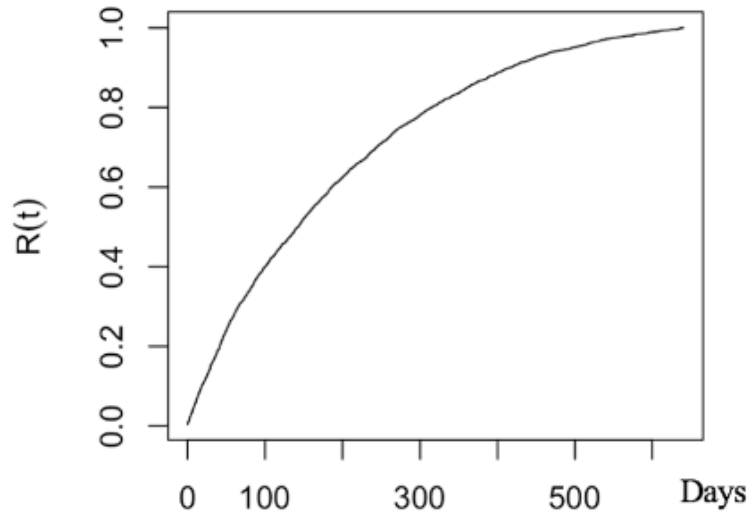


Figure 4.7: The rate of occurrence for the follow-up visit after treatment is initiated for all in-control data

4.6.2 Detect the Patients with Undesirable Outcomes

Furthermore, in the test dataset, there are 134 days signaled as out-of-control by the proposed monitoring method but not the traditional Chi-squared chart. To get further

insight into this, we study one of the 134 patients in further detail. Figure 4.8 shows the visit trajectory from 2008 to 2010 for this patient. We can observe that he/she initiated the AUD treatment on 2008-1-20, and his/her first follow-up visit occurred on 2009-4-30. The duration was over 15 months. Besides there was a 6-day hospitalization due to the excessive delay of the first follow-up visit with a physician. This is clearly an out-of-control patient with an undesired outcome whose behavior deviated from the ideal model, and we are able to detect it using the proposed method.

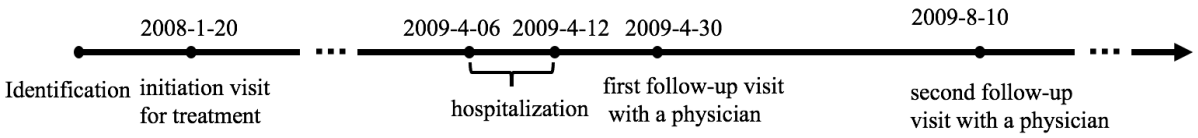


Figure 4.8: The visit trajectory for a patient who is identified as out-of-control by our proposed method but not the traditional method

4.7 Identify Key Factors Affecting the Patient Outcome

This section aims to develop an efficient classification model for patient outcome by comparing 8 machine learning approaches, such as logistic regression, Naive Bayes classification, support vector machines, linear discriminant analysis, k-nearest neighbors classification, decision tree classification, random forest classification and XGBoost classification. The patient outcomes are classified as two categories - in-control and out-of-control outcomes, where the patients with out of control outcome are identified by our proposed statistical monitoring method. The in-control outcome is referred as positive outcome, and out-of-control outcome is referred negative or undesirable outcome. In order to find the best classification method, the accuracy score will be used to measure the performance of the eight models. Based on the proposed best prediction model, the relationship between the various patient features and treatment outcome is identified, also the most influential factor on treatment outcome is

obtained. In the meantime, valuable insight and recommendation for improving the patient outcome is provided.

4.7.1 Machine Learning Methods

The section introduce the formulations and algorithms of the machine learning methods that we use. Given a dataset $\{y_i, x_{i1}, \dots, x_{ip}\}_{i=1}^{n}$ of n units, p stands for the number of predictor variables or features of each unit.

The first method is logistic regression. Logistic regression is a classification method that works for target value that is categorical. There are two steps for performing the logistic regression to predict the target with two possible outcomes, say 1 or 0. The first step is to perform a linear regression to build relationships between variables and then get an output P_i that shows the probability of the unit i belonging to the first class. The formula is shown below:

$$P_i(X) = \frac{1}{1 + \exp\left[-(b_0 + \sum_i^p b_i x_{ip})\right]}.$$

The threshold of the classification line is assumed to be at 0.5. For example, if the probability of one class I is greater than 0.5, we say the data is classified as class I. The involved parameters like b_0, b_i for $i \in [1, \dots, p]$ are estimated based on maximizing the likelihood function

$$L = \prod_{i=1}^{n} p(X_i)^{y_i}(1 - p(X_i))^{(1-y_i)}.$$

The second method is Naive Bayes classification. Naive Bayes classifier is one of the simplest and effective classification methods. It is based on Bayes' theorem to evaluate the probability of an event given a prior knowledge that related to the data. Let X denotes the predictors or features of an input variable and Y denotes a target variable with two possible outcomes. The mathematical formula is shown below:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)},$$

where $P(Y|X)$ is the value that we want to find, which denotes as the posterior probability of the class Y given X (a vector of the features for a unit). $P(X|Y)$ is the likelihood of X given class Y, which is calculated using the number of similar data points to X in the class Y divided by the number of total data points in the class Y. $P(Y)$ is the prior probability of the class Y, which is calculated using the number of data points in the class Y divided by the total number of data points. $P(X)$ is the prior probability or marginal likelihood of X, which is calculated using the number of similar data points to X divided by the total number of data points. A feature variable will be labeled as a class that has the highest posterior probability, which is typically greater than 0.5. For example, for a feature variable X with two possible class outcomes, say $Y = 1$ or $Y = 2$. If $P(Y = 1|X) > 0.5$, then X is determined to belong to class 1. If $P(Y = 1|X) < 0.5$, then X is determined to belong to class 2.

The third method is support vector machines. Support vector machines (SVM) is a method to define a decision boundary to separate the observed data into different classes. The decision boundary is actually a hyperplane, the objective of SVM is finding the optimal hyperplane that maximizes the margin between the two different classes, where margin is defined as Euclidean distance between the hyperplane and the closest point. The mathematical formula using to determine the boundary is shown below:

$$f(x) = \beta_0 + \sum \alpha_i K(x, x_i),$$

where α_i and β_0 are the training parameters and K denotes the kernel function. The decision boundary can be both linear and non-linear which depends on the kernel function we choose. The following are four typical kernel functions in performing SVM.

$$K(\mathbf{X_i}, \mathbf{X_j}) = \begin{cases} \mathbf{X_i} \cdot \mathbf{X_j} & \text{Linear} \\ (\gamma \mathbf{X_i} \cdot \mathbf{X_j} + \mathrm{C})^{\mathrm{d}} & \text{Polynomial} \\ \exp(-\gamma |\mathbf{X_i} - \mathbf{X_j}|^2) & \text{RBF} \\ \tanh(\gamma \mathbf{X_i} \cdot \mathbf{X_j} + \mathrm{C}) & \text{Sigmoid} \end{cases}$$

Figure 4.9: Typical kernel functions used in Support Vector Machines

Please note that the degree of the polynomial should be specified in polynomial kernel. And the radial basis function (RBF) kernel is adopted in sklearn by default, it usually performs good if feature variables have non-linear relationship. The sigmoid kernel is used for binary classification, which is similar to the concept of logistic regression.

The fourth method is linear discriminant analysis. Linear discriminant analysis (LDA) is a robust classification method with the assumptions that data follows multivariate Gaussian distribution and their covariance matrix are same among different classes. For a classification problem with K classes and N observations. The covariance matrix is defined as

$$\hat{\Sigma} = \sum_{k=1}^{K} \frac{1}{N-K} \sum_{i} (x_i - \hat{\mu_k})(x_i - \hat{\mu_k})^T,$$

where the mean of the class k is defined as

$$\hat{\mu_k} = \frac{1}{N_k} \sum_{i} x_i.$$

Then the computation steps are summarized in Figure 4.10. The $\hat{\pi}_k$ in the step 4 denotes the prior probability of class k, $\delta_k(x^*)$ is used to determine which class x belongs to. For example, for a classification problem with two classes k and l, We label x to class k if $\delta_k(x^*) - \delta_l(x^*) > 0$.

The fifth method is k-nearest neighbors (KNN). K-nearest neighbors classification method is a non-parametric classification method by identifying K nearest points to an observed point by measuring the distance. Then the observed data is classified to the class that appeared

1. Perform eigen-decompostion on the pooled covariance matrix: $\hat{\mathbf{\Sigma}} = \mathbf{U}\mathbf{D}\mathbf{U}^T$.
2. Sphere the data: $\mathbf{X}^* \leftarrow \mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T\mathbf{X}$.
3. Obtain class means in the transformed space: $\hat{\mu}_1, \ldots, \hat{\mu}_K$.
4. Classify $\mathbf{x}$ according to $\delta_k(\mathbf{x}^*)$:

$$\delta_k(\mathbf{x}^*) = \mathbf{x}^{*T}\hat{\mu}_k - \frac{1}{2}\hat{\mu}_k^T\hat{\mu}_k + \log\hat{\pi}_k \, .$$

Figure 4.10: The LDA computation steps

the most times among these K nearest points. The steps of KNN classification algorithm are summarized as follow:

1. Obtain a new unclassified data.
2. Measure the distance from the new data to all other data that are already labeled with a class. The distance can be measured using Euclidian, Manhattan, Minkowski or Weighted distance, where Euclidean distance is the most commonly used method to measure the distance between two points, its mathematical formulation is shown below:

$$d(X_1, X_2) = \sqrt{\sum_{j=1}^{p}(X_1 - X_2)^2},$$

   where p is the number of predictors or features of the data.
3. Get the K nearest points based on the distances.
4. Count the amount of each class in these K nearest points.
5. Label the new data to the class that appeared the most times among these K nearest points.

Please note that K is user specified parameter. Small K means the model is low bias and high variance, while large K will cause the model to become high bias and low variance. Thus,

it is important to choose a proper K to achieve the trade-off between bias and variance. In my dissertation, cross-validation method is used to select the best value of K.

The sixth method is decision tree classification. Decision tree is a classification method with the form of a tree structure. It divides the data into smaller subsets based on the feature variables in the dataset. The separation threshold of each decision node is the mean or mode of the respective feature variable. Entropy and information gain are used as criteria to split the data into child nodes or test the purity of the split. Entropy is defined as

$$E = -\sum_{k=1}^{K} p_k log(p_k),$$

which can be explained as the degree of uncertainty in the randomness of data. And the information gain is defined as

$$G = -\sum_{k=1}^{K} p_k(1 - p_k),$$

which measures the relative information contained by each feature. In both equations, p_k denotes the proportion of data that belongs to class k at each decision node. We split the data based on the feature with the minimum value of entropy or Gini index. The final result of decision tree classification looks like a tree that characterized with decision nodes and leaf nodes.

The seventh method is random forest classification. Random forest is an ensemble modeling technique that combines the output results across multiple individual decision trees. The decision tree algorithms are known for their simplicity and efficiency for dealing with dataset with large number of attributes. Beginning from the top of the tree, decision trees are generated by recursively splitting the training data to smaller subsets or regions based on the feature variables in the dataset. At each step of the splitting process, the best split is performed at a particular node without considering splits in future nodes. For each splitting process, information gain and mean squared error are commonly used as a criterion to select the best feature and determine the threshold for a splitting. Single decision tree tends to

over-fitting. However, the random forest technique can handle the over-fitting by building multiple decision trees using bootstrapped data from the training dataset and also selecting a random subset of the original features for splitting a node. The decision forest algorithm then estimates target value by averaging the predictions of the individual decision trees. The procedures of random forest are as follows:

Algorithm 4.1 Random Forest Classification

1. For b = 1 to B

   (i) Generate a bootstrap sample Z from the training data

   (ii) Grow a decision tree T_b to bootstrapped data Z by recursively selecting a random subset of the original features for splitting a node, until the maximum depth is reached

2. Output the ensemble trees $\{T_b\}_1^B$
3. Make a prediction to a new unclassified data point x: $\tilde{f}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$

The eighth method is XGBoost method. As another example of an ensemble model, boosting method is one of the most powerful learning model introduced in the last decades. Similar to random forest, the procedure of boosting is to combine the outputs of many "weak" learners to produce a powerful predictor. The general idea of boosting algorithms is to produce models sequentially, where each subsequent model attempts to correct the errors of its predecessor. Boosting method sequentially produces a series of weak learners, $\{G_m(x)\}_1^M$, to fit the data that has been modified repeatedly. Since weak learners are produced sequentially, the models cannot be parallel trained, because we must wait until the previous model has been trained and evaluated to generate the next model. The most commonly known boosting algorithm is AdaBoost. This method repeatedly corrects the learning model by paying more attention to training instances that were incorrectly predicted by the previous model.

Now instead of adjusting the instances weights at every iteration in AdaBoost, another more powerful boosting algorithm, gradient boosting, attempts to update the predicting

model with an additive form iteratively using gradient descent. The procedures of gradient boosting are as follows:

Algorithm 4.2 Gradient Boosting Method

1. Initial the boosted model on the original data, call it $F_0(x)$, by minimizing the loss function.

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

2. Compute the gradient of the loss function iteratively, which is same as the residual at each iteration. If we select square loss as the loss function,

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

3. Fit on the gradient obtained at each step and denote it as $h_m(x)$

4. Update the boosted model $F_m(x)$ as

$$F_m(x) = F_{m-1}(x) + \alpha h_m(x)$$

where α is the learning rate that is typically defined between [0,1].

5. Repeat step 2 to 4 until the loss is negligible, or the maximum limit of the number of estimators is reached

---

In addition to the classical gradient boosting algorithm, XGBoost is an advanced implementation of the Gradient Boosting. Thus XGBoost has the same learning procedures gradient boosting algorithm which described in Algorithm 4.2. This XGBoost algorithm has a stronger predicting power and faster speed than any other gradient boosting techniques. Specifically, XGBoost can control over-fitting by adding some regularization through both L1 and L2 penalization, handling sparse data, parallel learning, tree pruning, etc., which make sure XGBoost gives a better performance. Therefore, in this dissertation, we consider

using this powerful XGBoost boosting method, and the decision tree algorithm is used as the basis weak leaner in the boosting methods.

4.7.2 Model Performance Comparison

Based on the beneficiary summary DE-SynPUF data that is available on the website of the U.S. Centers for Medicare and Medicaid Services (*https://www.cms.gov*), we can get 5 attributes for each patient, including demographic, clinical and financial factors, which is shown in Figure 4.11. The attributes are comprised of continuous variables and categorical variables. The continuous variables include age, medical expenses, where medical expenses represent patient's out-of-pocket total medical expenses for inpatient, outpatient and carrier visits. The categorical variables include gender, race and if the patient has other chronic medical conditions. Note that the categorical variables need to be converted to dummy variables so that they can be used in classification models. Thus, each class of a categorical variable is treated as an attribute. For example, the race has 4 levels: White, Black, Hispanic, Others, then we can convert race into 3 dummy variables which are all denoted as binary variables. Finally, we will obtain total 7 attributes after converting all the categorical variables to dummy variables.

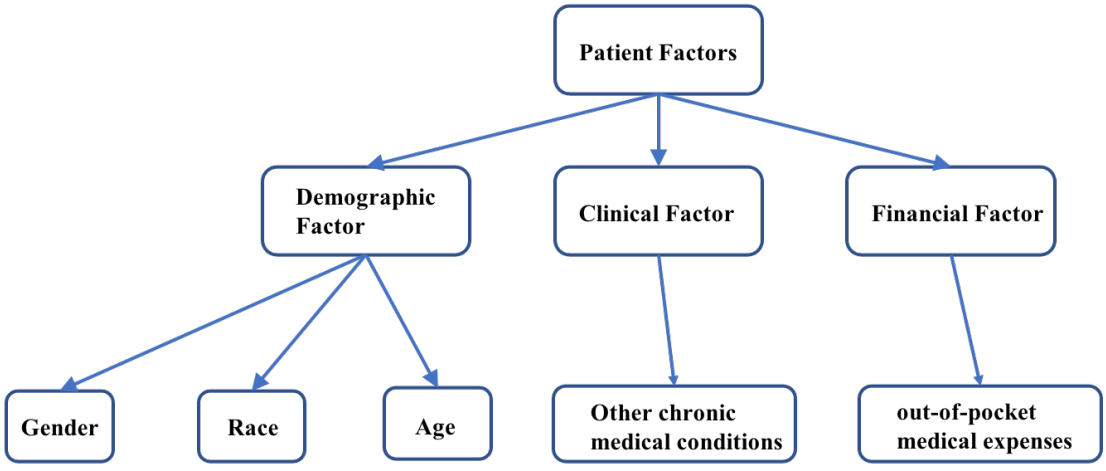


Figure 4.11: Patient factors that may affect the patient outcome

The correlation matrix between features is shown in Figure 4.12. The correlation coefficient is a measure of the degree of linear relationship between independent variables and dependent variable. The coefficient value can range from -1 to 1. The value of -1, 0, 1 indicate a perfect negative linear relationship, no linear relationship, and a perfect positive linear relationship between variables, respectively. Typically, if we find strong correlation between some features, some algorithms cannot deal with it very well and result in bad prediction performance. In that case, feature deduction methods such as principal component analysis can be used to reduce the dimension of the data and then overcome the drawback of strong correlation between these variables. However, in our data, as shown in Figure 4.12 , there is no strong correlation within different features, which means we do not need to worry about the accuracy of the classification methods would be affected and to perform the additional procedure to remove some features that are considered redundant.

In the subsection, in order to find the best classification method to build our treatment outcome prediction model for AUD patients, eight different classification techniques, such as logistic regression, Naive Bayes classification, support vector machines, linear discriminant analysis, k-nearest neighbors classification, decision tree classification, random forest classification, and XGBoost classification, have been compared and evaluated in terms of the prediction accuracy score (model score). In classification, the accuracy score is a statistical measure of how well the classification predictions match the real data class. Accuracy score is formulated as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}.$$

The accuracy can also be calculated using the confusion matrix, accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$, where TP is the true positives, TN is the true negatives, FP is the false positives, FN is the false negatives. Details of the model parameter settings in performance comparison analysis are summarized in table 4.1.

Table 4.1: Details of model parameter settings in performance comparison analysis

Model Parameter Settings		
Model	**Parameter Settings**	**Value**
Logistic Regression	Solver	lbfgs
	Tolerance for stopping criteria	1e-4
	Maximum number of iterations	100
	Penalty	L2
Naive Bayes	Data distribution assumption	Gaussian distribution
Support Vector Machines	Kernel function	Linear function
	Tolerance for stopping criteria	1e-3
	Regularization parameter	1
LDA	Solver	svd
	Tolerance for stopping criteria	1e-4
K-nearest Neighbors	Number of neighbors	9
	Distance metric	Euclidean metric
	Weights	Uniform
Decision Tree Classification	Tree splitting criterion	gini
	Minimum n. required to split	2
	Minimum node size	1
Random Forest Classification	Number of trees	100
	Tree splitting criterion	gini
	Minimum n. required to split	2
	Minimum node size	1
XGBoost Classification	Booster	Trees
	Number of trees	100
	Learning rate	0.1
	Maximum depth of a tree	3
	L2 regularization value	1

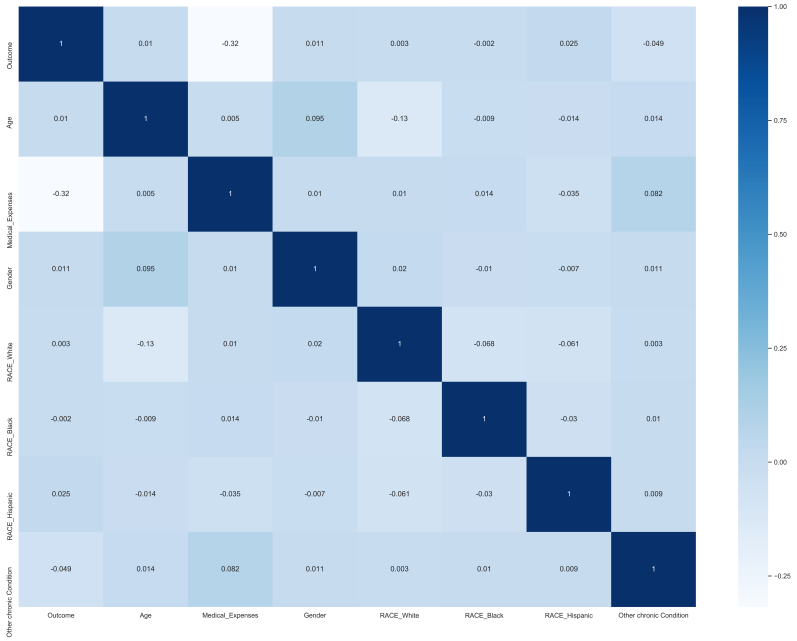


Figure 4.12: Correlation matrix between features

Table 4.2: Model performance comparison

Model	Accuracy scores
XGBoost Classification	0.876
Logistic Regression	0.860
Support Vector Machines	0.858
Linear Discriminant Analysis	0.854
K-nearest Neighbor Classification	0.816
Random Forest Classification	0.802
Decision Tree Classification	0.761
Naive Bayes Classification	0.643

Table 4.2 is the model performance comparison table. As shown in table 4.2, XGBoost gives us the highest accuracy score, which is 87.6 %. It indicates that XGBoost classification model delivers the best performance. On the other hand, the Naive Bayes classification has the worst performance because Naive Bayes classification assumes that all features are totally

independent which is impossible in real world. However XGBoost is an ensemble machine learning model based on a gradient boosting framework with various advanced enhancements to prevent overfitting. It can better integrate the outputs of many weak learners to produce a powerful predictor.

4.7.3 Feature Importance

Then, we examined the relative contribution of different patient characteristics. Figure 4.13 below summarizes the relative importance scores for the patient characteristics with respect to the accuracy of XGBoosting classification model which is the best performing model among all the eight models. The feature importance is calculated based on the average improvement in training accuracy gained when using the corresponding feature to split the data for a tree. A feature with the highest value of this metric means it is the most relevant or important feature for classifying an observation when compared to other features. The results of Figure 4.13 suggest that medical expenses, which is the financial factor, has the most significant impact on patient treatment outcome. And whether the patient has other chronic conditions has the least impact on patient treatment outcome.

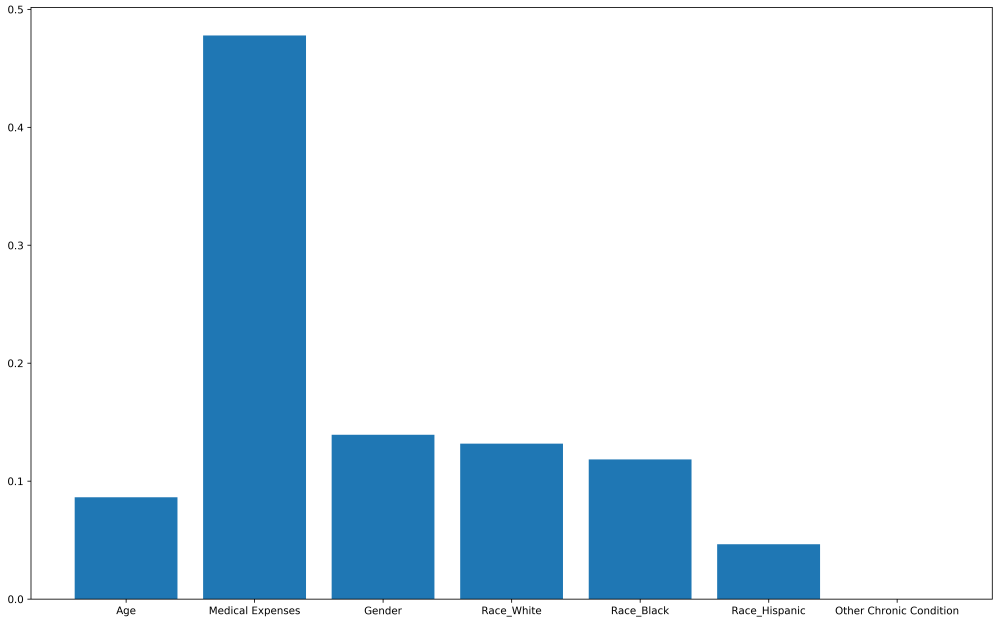


Figure 4.13: Patient's feature importances in affecting treatment outcome

To better understand the financial factor with respect to the treatment outcome. The boxplot of the out-of-pocket medical expenses for different treatment outcomes is shown in Figure 4.14 . We can see the median out-of-pocket medical expenses for patients with good outcome and undesirable outcome are $5412 and $8602, respectively. Also, the mean out-of-pocket medical expenses for patients with good outcome and undesirable outcome are $6172 and $9216, respectively. Thus, to improve the outcome, it is important to introduce financial incentive schemes for patients with AUD to encourage their continuity in care and adherence to the follow-up treatment. For example, patients can get reward if he/she initiates the treatment plan that made by the provider, and the reward is increasing over time as they continue to treatment. This kind of financial incentive program has achieved big success in improving the rate of follow-up treatment for other chronic medical conditions such as Opioid Use Disorder [120]. These financial incentive programs are usually supported by federal and state funding but have not been well utilized by promoting follow-up treatment for Alcohol Use Disorder.

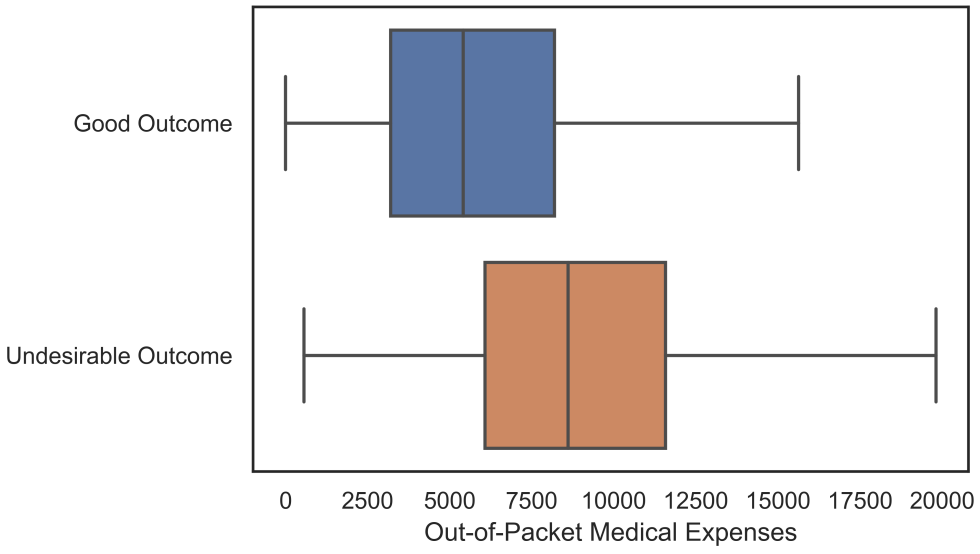


Figure 4.14: Boxplot of the out-of-pocket medical expenses

4.8 Conclusion

In this paper, we propose a continuous-time stochastic process model to measure and monitor the treatment process for patients with AUD based on the COC framework. First, the ideal rate of occurrence from one stage to another stage along COC can be estimated based on the in-control patients with favorable outcomes, from which benchmarks for COC can be developed. Furthermore, a new statistical monitoring scheme is developed to identify the patients whose care deviated from the baseline model, which would help clinicians develop subsequent interventions to improve outcomes. In the monitoring scheme, a test statistic $d^i_{p,q}(t)$ that measures deviation in the cumulative transition intensity is used to improve detection of performance. The effectiveness of the proposed method is demonstrated via simulations and a real case study, it is demonstrated that the proposed method outperforms the traditional Chi-squared chart. Furthermore, this chapter compares 8 different machine learning approaches to link the patient factors to adverse patient outcomes. It's demonstrated the XGBoost classification outperforms the other 7 classification models in terms of the evaluation of the model accuracy score. In the meantime, based on the training result of XGBoost classification model, we find that the financial factor is the key factor affecting treatment outcome. Thus, it is recommended to introduce financial incentive programs for patients with AUD to increase the rate of their follow-up treatment to improve outcome.

In the future study, extending the proposed monitoring method to detect COC events in the subpopulation level for AUD treatment is promising. In addition, in order to have a more accurate classification model to link the patient factors to treatment outcome, more features (including patient-level factors and facility-level factors) should be collected and considered in the future. For example, the counting processes $N^i_{p,q}(t)$ can be considered as covariates that are correlated to positive and negative treatment outcomes.

Chapter 5: Conclusion and Future Work

In this dissertation, a series of statistical monitoring methods based on stochastic process models are developed for detecting the abnormality in the timeliness of care and patient engagement for healthcare system to improve the quality of healthcare services. At the cute care service level, novel statistical monitoring approaches based on the log-likelihood ratio test and cumulative sum control chart are proposed to effectively detect the delay in service for emergency department that is modeled as different types of single queues and a network of queues, respectively. The developed statistical monitoring methods can be used as a retrospective evaluation tool. If a specific day for a node is signaled as out of control, the operations managers in the ED would retrospectively look into probable causes of the alarm and take necessary action to resolve it. For example, if we found a day of the week is signaled as out of control frequently, redesigning the weekly staffing schedule on that day could be considered. At the chronic care service level, a statistical monitoring scheme based on a continuous-time stochastic modeling approach is proposed for measuring and monitoring the Cascade of Care (COC) for patients with alcohol use disorder. The proposed model can identify the ideal patterns in the initiation and duration of AUD treatment for the key stages of the COC , from which benchmarks for COC can be developed. In the meantime, machine learning methods are applied to identify the key factor affecting treatment outcome, which will help inform the healthcare provider as well as public health associations to develop incentive programs to encourage treatment for patients with alcohol use disorder.

Chapter 2 focused on monitoring the timeliness of healthcare delivery in emergency department using counting processes. The proposed SQCT and GQCT methods are based on an approximate likelihood function that alleviates the issue of needing to numerically max-

imize a complex likelihood function for estimating the in-control parameters and obtaining test statistics. Both methods were shown to detect changes in the intensity that is otherwise hard to detect using existing techniques. Besides, the proposed scheme can be used to identify changes in real-time, which is particularly complicated for inhomogeneous queueing systems. The efficacy of the methods is demonstrated by simulation studies and a real-data case study. For future research, an optimal detection scheme that minimizes delay in change detection for inhomogeneous CTSPs would have a wide range of applications. Further theoretical development of the proposed GQCT method can lead to a better understanding of the type of penalty and the magnitude of the penalty that is ideal for detecting a specific kind of change in the intensity. Besides, monitoring the quality of care provided in other acute care sector such as ICU is an essential area of future research.

Chapter 3 focused on monitoring of the service rate in a network of queues with application in emergency department. Novel CUSUM control charts based on count data are proposed to monitor the service rate of a QN with time-inhomogeneous state dependent queues. The proposed CUSUM charts are compared with the MEWMA and MCUSUM charts using the ARL criteria to detect out-of-control scenarios. A major contribution of this research is the development of an easy to implement and efficient likelihood-ratio-based CUSUM charts, G-CUSUM and P-CUSUM charts for monitoring QNs, which could overcome the limitation of the normality assumption and do not need know the potential change in service rate of the queueing nodes in a QN, and thus have important practical applications. Numerical studies based on a simulated QN demonstrated that the proposed CUSUM charts can outperform traditional approaches on a variety of out-of-control scenario detection tests. Further, a case study focusing on monitoring the daily patient flow of an ED with multiple service stations demonstrates the efficacy of the proposed methods in a real application. For further research, extending our method to accommodate with the situation where only limited and partial samples can be observed is needed. Also, the study of approximation methods in establishing theoretical understanding of statistical monitoring

of QNs. Specifically, the application of diffusion approximation methods can help establish theoretical performance guarantees of CUSUM methods developed here.

Chapter 4 focused on monitoring and evaluating the Cascade of care for patients with alcohol use disorder. A novel statistical monitoring scheme with a continuous-time stochastic process model is proposed to monitor and measure the treatment process for patients with AUD based on the COC framework. First, the ideal rate of occurrence from one stage to another stage along COC can be estimated based on the in-control patients with favorable outcomes, from which benchmarks for COC can be developed. Furthermore, a new statistical monitoring scheme is developed to identify the patients whose care deviated from the baseline model, which is demonstrated to be superior than the conventional Chi-squared chart. In addition, various machine learning methods are adopted to investigate the relationship between patient factors and treatment outcome. Thus the key factor affecting treatment outcome can be identified, which would help clinicians or public health associations develop subsequent interventions to improve treatment outcomes for AUD. For further research, extending the proposed monitoring method to detect COC events in the subpopulation level for AUD treatment is promising. Also, more individual characteristics and other facility level characteristics should be considered to improve the classification accuracy thus more factors that led to negative outcomes can be identified.

References

[1] Kathleen N Lohr, Molla S Donaldson, and Jo Harris-Wehling. Medicare: a strategy for quality assurance, v: quality of care in a changing health care environment. *QRB. Quality review bulletin*, 18(4):120–126, 1992.

[2] Ali Mohammad Mosadeghrad. Healthcare service quality: towards a broad definition. *International journal of health care quality assurance*, 2013.

[3] Huilong Duan, Zhoujian Sun, Wei Dong, Kunlun He, and Zhengxing Huang. On clinical event prediction in patient treatment trajectory using longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*, 24(7):2053–2063, 2019.

[4] M Sulek Joanne, R Lin Mary, and S Marucheck Ann. Assessing the outcomes of quality improvement interventions. therole of x-chart methodology. *International Journal of Quality & Reliability Management*, 12(9):170–182, 1995.

[5] Uday M Apte and Charles C Reynolds. Quality management at kentucky fried chicken. *Interfaces*, 25(3):6–21, 1995.

[6] Ambreen Shafqat, Zhensheng Huang, and Muhammad Aslam. Design of x-bar control chart based on inverse rayleigh distribution under repetitive group sampling. *Ain Shams Engineering Journal*, 12(1):943–953, 2021.

[7] Joyce S Mehring. Achieving multiple timeliness goals for auto loans: a case for process control. *Interfaces*, 25(4):81–91, 1995.

[8] Rex S Green. The application of statistical process control to manage global client outcomes in behavioral healthcare. *Evaluation and Program Planning*, 22(2):199–210, 1999.

[9] AFB Costa and MA Rahim. A synthetic control chart for monitoring the process mean and variance. *Journal of Quality in Maintenance Engineering*, 12(1):81–88, 2006.

[10] Nurudeen Ayobami Ajadi, Osebekwin Asiribo, and Ganiyu Dawodu. Progressive mean exponentially weighted moving average control chart for monitoring the process location. *International Journal of Quality & Reliability Management*, 38(8):1680–1694, 2020.

[11] William H Woodall. The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, 38(2):89–104, 2006.

[12] John B Jensen and Robert E Markland. Improving the application of quality conformance tools in service firms. *Journal of Services Marketing*, 10(1):35–55, 1996.

[13] Francisco Aparisi, Charles W Champ, and J Carlos García-Díaz. A performance analysis of hotelling's $\chi 2$ control chart with supplementary runs rules. *Quality Engineering*, 16(3):359–368, 2004.

[14] VB Ghute and DT Shirke. A multivariate synthetic control chart for monitoring process mean vector. *Communications in Statistics—Theory and Methods*, 37(13):2136–2148, 2008.

[15] Pramod Dargopatil and Vikas Ghute. New sampling strategies to reduce the effect of autocorrelation on the synthetic t2 chart to monitor bivariate process. *Quality and Reliability Engineering International*, 35(1):30–46, 2019.

[16] Hengameh Hadian and Ali Rahimifard. Multivariate statistical control chart and process capability indices for simultaneous monitoring of project duration and cost. *Computers & Industrial Engineering*, 130:788–797, 2019.

[17] Khai Wah Khaw, Xinying Chew, Wai Chung Yeong, and Sok Li Lim. Optimal design of the synthetic control chart for monitoring the multivariate coefficient of variation. *Chemometrics and Intelligent Laboratory Systems*, 186:33–40, 2019.

[18] S Samanta and S Mondal. An application of multivariate control chart for online process monitoring in smes. In *Intelligent Electrical Systems: A Step towards Smarter Earth*, pages 191–198. CRC Press, 2021.

[19] Shuguang He, Wei Jiang, and Houtao Deng. A distance-based control chart for monitoring multivariate processes using support vector machines. *Annals of Operations Research*, 263(1):191–207, 2018.

[20] Rashid Mehmood, Muhammad Hisyam Lee, Iftikhar Ali, Muhammad Riaz, and Shahid Hussain. Multivariate cumulative sum control chart and measure of process capability based on bivariate ranked set schemes. *Computers & Industrial Engineering*, 150:106891, 2020.

[21] FuPeng Xie, JinSheng Sun, Philippe Castagliola, XueLong Hu, and Anan Tang. A multivariate cusum control chart for monitoring gumbel's bivariate exponential data. *Quality and Reliability Engineering International*, 37(1):10–33, 2021.

[22] Jean-Claude Malela-Majika, Kashinath Chatterjee, and Christos Koukouvinos. A multivariate triple exponentially weighted moving average control chart. *Quality and Reliability Engineering International*, pages 1–32, 2021.

[23] Jimoh Olawale Ajadi, Inez Maria Zwetsloot, and Kwok-Leung Tsui. A new robust multivariate ewma dispersion control chart for individual observations. *Mathematics*, 9(9):1038, 2021.

[24] Patrick D Bourke. The geometric cusum chart with sampling inspection for monitoring fraction defective. *Journal of Applied Statistics*, 28(8):951–972, 2001.

[25] Zhang Wu, Song Huat Yeo, and Trevor A Spedding. A synthetic control chart for detecting fraction nonconforming increases. *Journal of Quality Technology*, 33(1):104–111, 2001.

[26] MP Gadre and RN Rattihalli. Unit and group-runs chart to identify increases in fraction nonconforming. *Journal of quality technology*, 37(3):199–209, 2005.

[27] Jian Li, Fugee Tsung, and Changliang Zou. A simple categorical chart for detecting location shifts with ordinal information. *International journal of production research*, 52(2):550–562, 2014.

[28] Jiayun Jin and Geert Loosveldt. Nonparametric multivariate control chart for numerical and categorical variables. *Communications in Statistics-Simulation and Computation*, pages 1–19, 2021.

[29] Joanne M Sulek, Ann Marucheck, and Mary R Lind. Measuring performance in multistage service operations: An application of cause selecting control charts. *Journal of Operations Management*, 24(5):711–727, 2006.

[30] Katina R Skinner, Douglas C Montgomery, and George C Runger. Process monitoring for multiple count data using generalized linear model-based control charts. *International Journal of Production Research*, 41(6):1167–1180, 2003.

[31] Fatemeh Sogandi, Majid Aminnayeri, Adel Mohammadpour, and Amirhossein Amiri. Phase i risk-adjusted bernoulli chart in multistage healthcare processes based on the state-space model. *Journal of Statistical Computation and Simulation*, 91(3):522–542, 2021.

[32] Jon Mark Hirshon, Nicholas Risko, Emilie JB Calvello, Sarah Stewart de Ramirez, Mayur Narayan, Christian Theodosis, and Joseph O'Neill. Health systems and services: the role of acute care. *Bulletin of the World Health Organization*, 91:386–388, 2013.

[33] Caroline A Brand, Melinda Martin-Khan, Olivia Wright, Richard N Jones, John N Morris, Catherine M Travers, Joannne Tropea, and Leonard C Gray. Development of quality indicators for monitoring outcomes of frail elderly hospitalised in acute care health settings: study protocol. *BMC Health Services Research*, 11(1):1–8, 2011.

[34] Greet Baldewijns, Stijn Luca, William Nagels, Bart Vanrumste, and Tom Croonenborghs. Automatic detection of health changes using statistical process control techniques on measured transfer times of elderly. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5046–5049. IEEE, 2015.

[35] Ahsan H Khandoker, Daniel TH Lai, Rezaul K Begg, and Marimuthu Palaniswami. Wavelet-based feature extraction for support vector machines for screening balance impairments in the elderly. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 15(4):587–597, 2007.

[36] Ekachai Thammasat and Jarree Chaicharn. An application of a cumulative-sum control chart for elderly fall detection using smartphone accelerometers. *Science & Technology Asia*, pages 36–46, 2020.

[37] Anette H Ranhoff, AU Gjoen, M Mowe, et al. Screening for malnutrition in elderly acute medical patients: the usefulness of mna-sf. *J Nutr Health Aging*, 9(4):221–5, 2005.

[38] Nor Hasliza Mat Desa, Abdul Aziz Jemain, and Maznah Mat Kasim. Residual control chart for monitoring pediatrics hospital admission performances. *Contemporary Engineering Sciences*, 8(32):1509–1515, 2015.

[39] Margaret Hsiau, Hilda E Fernandez, David Gjertson, Robert B Ettenger, and Eileen W Tsai. Monitoring nonadherence and acute rejection with variation in blood immunosuppressant levels in pediatric renal transplantation. *Transplantation*, 92(8):918–922, 2011.

[40] David D'Arienzo, Erin Hessey, Rami Ali, Sylvie Perreault, Susan Samuel, Louise Roy, Jacques Lacroix, Philippe Jouvet, Genevieve Morissette, Marc Dorais, et al. A validation study of administrative health care data to detect acute kidney injury in the pediatric intensive care unit. *Canadian journal of kidney health and disease*, 6:2054358119827525, 2019.

[41] Travis J Moss, Matthew T Clark, James Forrest Calland, Kyle B Enfield, John D Voss, Douglas E Lake, and J Randall Moorman. Cardiorespiratory dynamics measured from continuous ecg monitoring improves detection of deterioration in acute care patients: A retrospective cohort study. *PLoS One*, 12(8):e0181448, 2017.

[42] David A Cook. *The development of risk adjusted control charts and machine learning models to monitor the mortality rate of intensive care unit patients*. University of Queensland, 2003.

[43] David A Cook, Stefan H Steiner, Richard J Cook, Vern T Farewell, and Anthony P Morton. Monitoring the evolutionary process of quality: risk-adjusted charting to track outcomes in intensive care. *Critical care medicine*, 31(6):1676–1682, 2003.

[44] Antonie Koetsier, Nicolette F de Keizer, Evert de Jonge, David A Cook, and Niels Peek. Performance of risk-adjusted control charts to monitor in-hospital mortality of intensive care unit patients: a simulation study. *Critical care medicine*, 40(6):1799–1807, 2012.

[45] Isabela Pereira Rodrigues, Osiris Turnes, and Celeste Aida Nogueira Silveira. Surveillance analysis and monitoring of mutidrug-resistant bacteria incidence in an intensive care unit: the role of cumulative sum control charts. *International Journal of Sciences: Basic and Applied Research*, 23(1):217–229, 2015.

[46] Christine S Cocanour, Michelle Peninger, Bradley D Domonoske, Tao Li, Bobbie Wright, Alicia Valdivia, and Katharine M Luther. Decreasing ventilator-associated pneumonia in a trauma icu. *Journal of Trauma and Acute Care Surgery*, 61(1):122–130, 2006.

[47] Stephanie Medlock, Saeid Eslami, Marjan Askari, Erik Jan van Lieshout, Dave A Dongelmans, and Ameen Abu-Hanna. Improved communication in post-icu care by improving writing of icu discharge letters: a longitudinal before-after study. *BMJ quality & safety*, 20(11):967–973, 2011.

[48] M A Mohammed. Using statistical process control to improve the quality of health care. *BMJ Quality & Safety*, 13(4):243–245, 2004.

[49] Farid Kadri, Fouzi Harrou, Sondès Chaabane, Ying Sun, and Christian Tahon. Seasonal arma-based spc charts for anomaly detection: Application to emergency department systems. *Neurocomputing*, 173:2102–2114, 2016.

[50] Frank C Kaminsky, John Maleyeff, Sherry Providence, Esther Purinton, and Mary Waryasz. Using spc to analyze quality indicators in a healthcare organization. *Journal of Healthcare Risk Management*, 17(4):14–22, 1997.

[51] Charles D Callahan and David L Griffen. Advanced statistics: applying statistical process control techniques to emergency medicine: a primer for providers. *Academic emergency medicine*, 10(8):883–890, 2003.

[52] Hugh Rogers, Stephen Gilligan, and Melanie Walters. Quality improvements in hospital flow may lead to a reduction in mortality. *Clinical Governance: An International Journal*, 2008.

[53] Christina Pagel, Padmanabhan Ramnarayan, Samiran Ray, and Mark J Peters. Development and implementation of a real time statistical control method to identify the start and end of the winter surge in demand for paediatric intensive care. *European Journal of Operational Research*, 264(3):847–858, 2018.

[54] John L Moran and Patricia J Solomon. Statistical process control of mortality series in the australian and new zealand intensive care society (anzics) adult patient database: implications of the data generating process. *BMC medical research methodology*, 13(1):66, 2013.

[55] Nan Chen and Shiyu Zhou. Cusum statistical monitoring of m/m/1 queues and extensions. *Technometrics*, 57(2):245–256, 2015.

[56] Chén C Kenyon, David A Hill, Sarah E Henrickson, Tyra C Bryant-Stephens, and Joseph J Zorc. Initial effects of the covid-19 pandemic on pediatric asthma emergency department utilization. *The Journal of Allergy and Clinical Immunology: In Practice*, 8(8):2774–2776, 2020.

[57] Fouzi Harrou, Farid Kadri, Sondes Chaabane, Christian Tahon, and Ying Sun. Improved principal component analysis for anomaly detection: Application to an emergency department. *Computers & Industrial Engineering*, 88:63–77, 2015.

[58] Olatunde A Adeoti. Application of cusum control chart for monitoring hiv/aids patients in nigeria. *International Journal of Statistics and Applications*, 3(3):77–80, 2013.

[59] Robin M Turner, Andrew Hayen, Petra Macaskill, Les Irwig, and Helen K Reddel. Control charts demonstrated limited utility for the monitoring of lung function in asthma. *Journal of clinical epidemiology*, 65(1):53–61, 2012.

[60] Fazel Hayati, Seed Maghsoodloo, Michael J DeVivo, and Brian J Carnahan. Control chart for monitoring occupational asthma. *Journal of Safety Research*, 37(1):17–26, 2006.

[61] Pershang Dokouhaki and Rassoul Noorossana. Surveillance of diabetes prevalence rate through the development of a markov-based control chart. *Journal of Mechanics in Medicine and Biology*, 12(04):1250083, 2012.

[62] Muhammad Aslam, Gadde Srinivasa Rao, Nasrullah Khan, and Fahad A Al-Abbasi. Ewma control chart using repetitive sampling for monitoring blood glucose levels in type-ii diabetes patients. *Symmetry*, 11(1):57, 2019.

[63] Katarzyna Kaczmarek-Majer, Olgierd Hryniewicz, Karol R Opara, Weronika Radziszewska, Anna Olwert, Jan W Owsiński, and Sławomir Zadrożny. Control charts designed using model averaging approach for phase change detection in bipolar disorder. In *International Conference Series on Soft Methods in Probability and Statistics*, pages 115–123. Springer, 2018.

[64] Casey B Cottrill, Stephanie Lemle, Steven C Matson, Andrea E Bonny, and Erin R McKnight. Multifaceted quality improvement initiative improves retention in treatment for youth with opioid use disorder. *Pediatric quality & safety*, 4(3), 2019.

[65] Murtadha Aldeer, Mehdi Javanmard, and Richard P Martin. A review of medication adherence monitoring technologies. *Applied System Innovation*, 1(2):14, 2018.

[66] Robert H Remien, Michael J Stirratt, Curtis Dolezal, Joanna S Dognin, Glenn J Wagner, Alex Carballo-Dieguez, Nabila El-Bassel, and Tiffany M Jung. Couple-focused support to improve hiv medication adherence: a randomized controlled trial. *Aids*, 19(8):807–814, 2005.

[67] Jonathan Hatoun, Megan Bair-Merritt, Howard Cabral, and James Moses. Increasing medication possession at discharge for patients with asthma: the meds-in-hand project. *Pediatrics*, 137(3), 2016.

[68] Sachini N Bandara, Hillary Samples, Rosa M Crum, and Brendan Saloner. Is screening and intervention associated with treatment receipt among individuals with alcohol use disorder? evidence from a national survey. *Journal of substance abuse treatment*, 92:85–90, 2018.

[69] Jessica E Haberer, Josh Kahane, Isaac Kigozi, Nneka Emenyonu, Peter Hunt, Jeffrey Martin, and David R Bangsberg. Real-time adherence monitoring for hiv antiretroviral therapy. *AIDS and Behavior*, 14(6):1340–1346, 2010.

[70] John G McHutchison, Michael Manns, Keyur Patel, Thierry Poynard, Karen L Lindsay, Christian Trepo, Jules Dienstag, William M Lee, Carmen Mak, Jean-Jacques Garaud, et al. Adherence to combination therapy enhances sustained response in genotype-1–infected patients with chronic hepatitis c. *Gastroenterology*, 123(4):1061–1069, 2002.

[71] Aldona Kubica, Agata Kosobucka, Piotr Michalski, Łukasz Pietrzykowski, Aleksandra Jurek, Marzena Wawrzyniak, and Michał Kasprzak. The adherence in chronic diseases scale—a new tool to monitor implementation of a treatment plan. *Folia Cardiol*, 12(1):19–26, 2017.

[72] Deborah Matteliano, Barbara J St Marie, June Oliver, and Candace Coggins. Adherence monitoring with chronic opioid therapy for persistent pain: a biopsychosocial-spiritual approach to mitigate risk. *Pain Management Nursing*, 15(1):391–405, 2014.

[73] Laxmaiah Manchikanti, Rajeev Manchukonda, Kim S Damron, Doris Brandon, Carla D McManus, and Kim Cash. Does adherence monitoring reduce controlled substance abuse in chronic pain patients? *Pain physician*, 9(1):57–60, 2006.

[74] William H Woodall. Control charts based on attribute data: bibliography and review. *Journal of quality technology*, 29(2):172–183, 1997.

[75] Sheldon M Ross. *Introduction to probability models*. Academic press, 2014.

[76] Patricia Reynaud-Bouret. Penalized projection estimators of the aalen multiplicative intensity. *Bernoulli*, 12(4):633–661, 2006.

[77] Thomas H Scheike. Additive–multiplicative intensity models. *Wiley StatsRef: Statistics Reference Online*, 2014.

[78] Patricia Reynaud-Bouret and Sophie Schbath. Adaptive estimation for hawkes processes; application to genome analysis. *The Annals of Statistics*, 38(5):2781–2822, 2010.

[79] Niels Richard Hansen, Patricia Reynaud-Bouret, and Vincent Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

[80] Stéphane Gaïffas and Agathe Guilloux. High-dimensional additive hazards models and the lasso. *Electronic Journal of Statistics*, 6:522–546, 2012.

[81] Sophie Donnet, Vincent Rivoirard, Judith Rousseau, and Catia Scricciolo. Posterior concentration rates for counting processes with aalen multiplicative intensities. *Bayesian Analysis*, 12(1):53–87, 2017.

[82] Per K Andersen, Ornulf Borgan, Richard D Gill, and Niels Keiding. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

[83] Changliang Zou, Xianghui Ning, and Fugee Tsung. Lasso-based multivariate linear profile monitoring. *Annals of Operations Research*, 192(1):3–19, 2012.

[84] Mustafa Y Sir, David Nestler, Thomas Hellmich, Devashish Das, Michael J Laughlin Jr, Michon C Dohlman, and Kalyan Pasupathy. Optimization of multidisciplinary staffing improves patient experiences at the mayo clinic. *Interfaces*, 47(5):425–441, 2017.

[85] Martin Reiser and Hisashi Kobayashi. Queuing networks with multiple closed chains: theory and computational algorithms. *IBM journal of Research and Development*, 19(3):283–294, 1975.

[86] Ronald G Askin and Girish Jampani Hanumantha. Queueing network models for analysis of nonstationary manufacturing systems. *International Journal of Production Research*, 56(1-2):22–42, 2018.

[87] Daniel A Menascé and Shouvik Bardhan. Tdqn: Trace-driven analytic queuing network modeling of computer systems. *Journal of Systems and Software*, 147:162–171, 2019.

[88] Satoshi Hoshino, Jun Ota, Akiko Shinozaki, and Hideki Hashimoto. Optimal design methodology for an agv transportation system by using the queuing network theory. In *Distributed Autonomous Robotic Systems 6*, pages 411–420. Springer, 2007.

[89] Na Li, Nan Kong, Quanlin Li, and Zhibin Jiang. Evaluation of reverse referral partnership in a tiered hospital system–a queuing-based approach. *International Journal of Production Research*, 55(19):5647–5663, 2017.

[90] Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, and Galit B Yom-Tov. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems*, 5(1):146–194, 2015.

[91] Reetu Mehandiratta. Applications of queuing theory in health care. *International Journal of Computing and Business Research*, 2(2):2229–6166, 2011.

[92] Jeffery K Cochran and Kevin T Roche. A multi-class queuing network analysis methodology for improving hospital emergency department performance. *Computers & Operations Research*, 36(5):1497–1512, 2009.

[93] Hajnal Vass and Zsuzsanna K Szabo. Application of queuing model to patient flow in emergency department. case study. *Procedia Economics and Finance*, 32:479–487, 2015.

[94] Junfei Huang, Boaz Carmeli, and Avishai Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

[95] Jingui Xie, Ping Cao, Boray Huang, and Marcus Eng Hock Ong. Determining the conditions for reverse triage in emergency medical services using queuing theory. *International Journal of Production Research*, 54(11):3347–3364, 2016.

[96] Pengyi Shi, Jonathan E Helm, H Sebastian Heese, and Alice M Mitchell. An operational framework for the adoption and integration of new diagnostic tests. *Production and Operations Management*, 30(2):330–354, 2021.

[97] William H Woodall, Benjamin M Adams, and James C Benneyan. The use of control charts in healthcare. *Statistical methods in healthcare*, 19:251–267, 2012.

[98] Kimberly D Johnson and Chris Winkelman. The effect of emergency department crowding on patient outcomes: a literature review. *Advanced emergency nursing journal*, 33(1):39–54, 2011.

[99] Jing Wen, Na Geng, and Xiaolan Xie. Real-time scheduling of semi-urgent patients under waiting time targets. *International Journal of Production Research*, 58(4):1127–1143, 2020.

[100] Changliang Zou and Fugee Tsung. Directional mewma schemes for multistage process monitoring and diagnosis. *Journal of Quality Technology*, 40(4):407–427, 2008.

[101] Martin I Reiman. The heavy traffic diffusion approximation for sojourn times in jackson networks. In *Applied probability—computer science: the interface*, pages 409–421. Springer, 1982.

[102] Jin Zhang. Powerful goodness-of-fit tests based on the likelihood ratio. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(2):281–294, 2002.

[103] Sumi Kim and Seongmoon Kim. Differentiated waiting time management according to patient class in an emergency care center using an open jackson network integrated with pooling and prioritizing. *Annals of Operations Research*, 230(1):35–55, 2015.

[104] Nan Chen, Yuan Yuan, and Shiyu Zhou. Performance analysis of queue length monitoring of m/g/1 systems. *Naval Research Logistics (NRL)*, 58(8):782–794, 2011.

[105] Sven Knoth. The art of evaluating monitoring schemes - how to measure the performance of control charts? In *Frontiers in statistical quality control 8*, pages 74–99. Springer, 2006.

[106] Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

[107] Ross S Sparks. Cusum charts for signalling varying location shifts. *Journal of Quality Technology*, 32(2):157–171, 2000.

[108] Muhammad Faisal, Raja Fawad Zafar, Nasir Abbas, Muhammad Riaz, and Tahir Mahmood. A modified cusum control chart for monitoring industrial processes. *Quality and Reliability Engineering International*, 34(6):1045–1058, 2018.

[109] Changliang Zou and Peihua Qiu. Multivariate statistical process control using lasso. *Journal of the American Statistical Association*, 104(488):1586–1596, 2009.

[110] Joseph J Pignatiello Jr and George C Runger. Comparisons of multivariate cusum charts. *Journal of quality technology*, 22(3):173–186, 1990.

[111] Cynthia A Lowry, William H Woodall, Charles W Champ, and Steven E Rigdon. A multivariate exponentially weighted moving average control chart. *Technometrics*, 34(1):46–53, 1992.

[112] Haim Shore. Control charts for the queue length in a g/g/s system. *IIE Transactions*, 38(12):1117–1130, 2006.

[113] Changliang Zou and Fugee Tsung. A multivariate sign ewma control chart. *Technometrics*, 53(1):84–97, 2011.

[114] Robert J Batt and Christian Terwiesch. Doctors under load: An empirical study of state-dependent service times in emergency care. *The Wharton School, the Univ ersity of Pennsylvania, Philadelphia, PA*, 19104, 2012.

[115] Yanqing Kuang, Devashish Das, Kimberly Johnson, and Mingyang Li. A continuous-time stochastic modeling approach for monitoring the cascade of care for patients with alcohol use disorder. In *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), 2022.

[116] Aaron M White, Megan E Slater, Grace Ng, Ralph Hingson, and Rosalind Breslow. Trends in alcohol-related emergency department visits in the united states: results from the nationwide emergency department sample, 2006 to 2014. *Alcoholism: clinical and experimental research*, 42(2):352–359, 2018.

[117] Lewei A Lin, Erin E Bonar, Lan Zhang, Rachel Girard, and Lara N Coughlin. Alcohol-involved overdose deaths in us veterans. *Drug and alcohol dependence*, 230:109196, 2022.

[118] K Witkiewitz, RZ Litten, and L Leggio. Advances in the science and treatment of alcohol use disorder. *Science advances*, 5(9):eaax4043, 2019.

[119] Ramnath Subbaraman, Ruvandhi R Nathavitharana, Kenneth H Mayer, Srinath Satyanarayana, Vineet K Chadha, Nimalan Arinaminpathy, and Madhukar Pai. Constructing care cascades for active tuberculosis: a strategy for program monitoring and identifying gaps in quality of care. *PLoS medicine*, 16(2):e1002754, 2019.

[120] Austin S Kilaru, Jeanmarie Perrone, David Kelley, Sari Siegel, Su Fen Lubitz, Nandita Mitra, and Zachary F Meisel. Participation in a hospital incentive program for follow-up treatment for opioid use disorder. *JAMA Network Open*, 3(1):e1918511–e1918511, 2020.

Appendix A: Copyright Permission

The permission below is for the reproduction of material in Chapter 4.

RE: Copyright Permission for Dissertation

Anna Johnston <ajohnston@iise.org>
Wed 3/30/2022 8:48 AM
To: Yanqing Kuang <ykuang@usf.edu>

Good morning.

IISE does grant you permission rights to reuse content as requested in your dissertation only. Please include a credit line in your dissertation.

Reproduced by permission for dissertation use only, "Title of Paper" in Proceedings of the 2022 Annual Conference, Institute of Industrial and Systems Engineers.

For any other use beyond your dissertation, please submit IISE copyright permission requests to the Copyright Clearance Center (CCC) at www.copyright.com, or toll-free: 1-855-239-3415 or by email: info@copyright.com.

Anna Johnston
Conference Manager | Institute of Industrial & Systems Engineers
(770) 349-1114 | ajohnston@iise.org
www.iise.org

What happens when a fleet safety expert, a Michelin star, an astronaut and an Olympian walk into keynote speaking slots? Your ergo programs benefit. Experience all-encompassing learning and networking at the Applied Ergonomics Conference March 21-24 in Orlando. Register & save today. Digital option available.

From: Yanqing Kuang <ykuang@usf.edu>
Sent: Tuesday, March 29, 2022 10:38 AM
To: Anna Johnston <ajohnston@iise.org>
Subject: Copyright Permission for Dissertation

Dear Anna,

This is Yanqing Kuang, a PhD student at the University of South Florida and an IISE member. One of my papers has been recently accepted for publication in the 2022 IISE Annual Conference. I'd like to ask for permission to use this paper as part of my PhD dissertation. The paper title is **A Continuous-time Stochastic Modeling Approach for Monitoring the Cascade of Care for Patients with Alcohol Use Disorder**.

I appreciate your time and consideration.

Best regards,
Yanqing

Appendix B: Supplemental Materials

B.1 Appendix for Chapter 2

B.1.1 Penalized GLR for Poisson Process

Considering the problem of identifying the decrease in the intensity of a Poisson process. The traditional GLR test (GLRT) for the hypothesis test $H_0 : \lambda = \lambda_0$ versus $H_0 : \lambda \neq \lambda_0$, for the in-control intensity λ_0 is

$$\max_{\lambda} \quad n \log\left(\frac{\lambda}{\lambda_0}\right) - \lambda T + \lambda_0 T \tag{B.1}$$

where n events were recorded in a $[0, T]$ time period. The two alternative penalization term result in two penalized GLR testing methods. They are referred to as PGLR1 and PGLR2, where

$$PGLR1 = \max_{\lambda} \quad n \log\left(\frac{\lambda}{\lambda_0}\right) - \lambda T + \lambda_0 T - (\psi T)(\lambda - \lambda_0)^2$$

$$PGLR2 = \max_{\lambda} \quad n \log\left(\frac{\lambda}{\lambda_0}\right) - \lambda T + \lambda_0 T - (\psi T)(\lambda)^2$$

The penalty (ψT) is assumed to be proportional to T. In order to compare the performance of these three tests, we use an example with $\lambda_0 = 1$ and $T = 1$. Figure B.1 shows that PGLRT2 with $\psi = 1$ has the best performance in detecting the decrease in λ for a simple Poisson process, which demonstrates the efficiency for the penalized GLR test by penalizing λ instead of $\lambda - \lambda_0$.

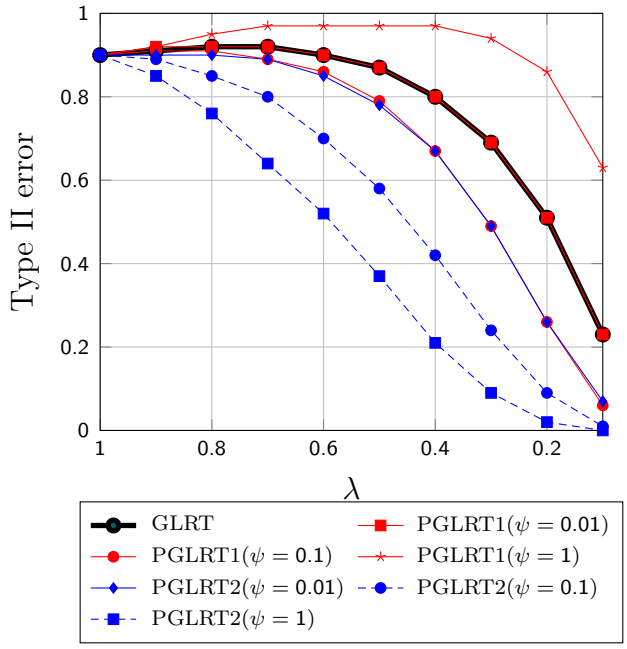


Figure B.1: Detecting decrease in λ for a simple Possion process using loglikelihood ratio test