PHASE I MONITORING WITH APPLICATIONS
IN MANUFACTURING AND
HEALTHCARE

by

SMRITI NEOGI

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
MAY 2014

Acknowledgements

I sincerely thank my advisor Dr. Li Zeng for accepting me as her student and patiently guiding me through this journey of research. She has taught me how to think instead of what to think. PhD is a test of patience and persistence and I am thankful for her trust and confidence in me which kept me going in the dark days when I lost faith in myself.

I am extremely thankful to Dr. Victoria Chen for introducing me to Dr. Zeng. It was difficult to accomplish my goals without their support and guidance. I wish to thank Dr Jamie Rogers, Dr. Donald Liles and Dr. Victoria Chen for their interest in my research and time to serve in dissertation committee I wish to thank Dr. Jamie Rogers and doctors of the University of Texas Medical Branch (UTMB) Galveston for providing us the healthcare data

I would like to thank the Graduate School and the Financial Aid Office for providing me the tuition support through the STEM scholarship. I thank my friends and COSMOS students Na Wang, Zirun Zhang, John Dickson, Alireza Norrozi, Aera Kim Leboullec and Kin Ming Kam for their unconditional support, the light moments we shared in our COSMOS lab and the nerdy discussions that enabled strong understanding of the subjects

I thank my loving husband Amaresh Satpathy for his unconditional love by motivating me at every step of this journey and support by staying separate that helped me stay focused on my goal. Last but not least, I thank and dedicate this work to my parents Mrs Jharna Neogi and Mr Swaraj Kumar Neogi, and my sister Ms Shrishti Neogi for their strong fight against the society to educate a girl child and for their ever increasing love and affection to me.

April, 16, 2014

Abstract

PHASE I MONITORING WITH APPLICATIONS IN MANUFACTURING AND HEALTHCARE

Smriti Neogi PhD

The University of Texas at Arlington, 2014

Supervising Professor: Li Zeng

This research develops statistical methods for quality monitoring in complex systems. Quality monitoring typically consists of two phases called Phase I analysis (or offline monitoring) and Phase II analysis (or online monitoring). This research is focused on Phase I monitoring. Two application areas are considered, complex manufacturing processes and healthcare delivery processes.

In the first application, a robust strategy for Phase I analysis of optical profiles in low-E glass manufacturing is developed. The proposed approach aims to solve the problems such as violation of normality, high dimensionality, detection of multiple change points, etc. It will provide a convenient process monitoring tool for practitioners in the low-E glass industry.

In the second application, a systematic methodology for Phase I monitoring of patient readmission is developed. Patient readmission is a critical contributor to the rising health care costs and has become an important performance indicator for assessing and monitoring quality of care. This work consists of two parts: construction of readmission model and change detection based on the constructed model. The proposed approach is demonstrated using real data from chronic obstructive pulmonary diseases (COPD) patients.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

Introduction

1.1 Motivation

Statistical Process Control (SPC) has become an integral part of continuous improvement (Montgomery, 2009). There are two main reasons: the economic downturn situation and the availability of abundant data on account of fast paced electronic data acquisition system. Companies are striving hard to survive in this economic downturn. In such a scenario delivering good-quality products/services through reducing process variability and defects has become essential than ever before. Another reason that emphasizes the importance of SPC is the availability of big data (Manyika et al., 2008). The advanced measurement/sensing technologies used today capture tons of data which can provide valuable information on the process which generated the data. Using those data, changes in the product/service quality can be detected and root causes of the changes can be found. However, one challenge in using the data is that some data are difficult to deal with using conventional SPC methods. For example, traditional process control methods are based on assumptions which may be violated in some processes due to the complex variation sources existing in the process. To conduct process control and quality improvement, it is very important to develop methods which can conquer the intrinsic complexity of the data. To fill in the gap in the literature, this dissertation aims to explore statistical methods for quality monitoring in complex systems. Two application areas are considered, complex manufacturing processes and healthcare delivery processes. More specifically, this research focuses on Phase I quality monitoring, a brief introduction on which is provided as follows.

Quality monitoring typically consists of two phases called Phase I analysis (or offline monitoring) and Phase II analysis (or online monitoring) (Sullivan, 2002). Figure

1.1 shows a schematic overview of the two types of monitoring. Basically, Phase I monitoring is used to detect changes in the historical data. The purpose is to identify data from the in-control condition and estimate the parameters of the in-control model. Such estimates will then be used to establish the monitoring system. Once the monitoring system becomes available, it will be used to inspect online data to determine if the process is in control or out of control, which is the main task of Phase II monitoring. Usually, when a change is detected, the process will be stopped and root causes of the change will be identified. Adjustment will then be made in the process to fix the problems to bring the process back to normal.

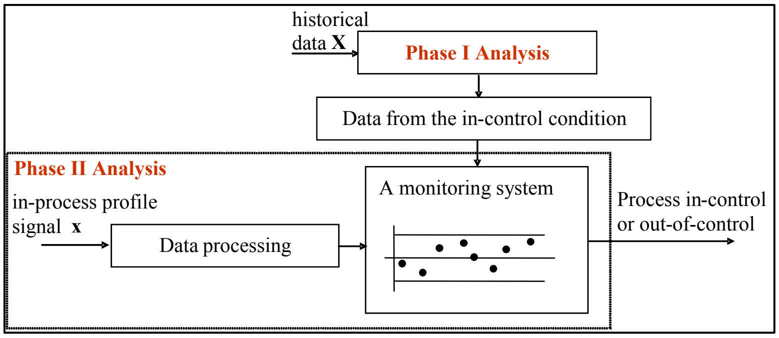


Figure 1-1 Overview of Phase I and Phase II Monitoring

Figure 1.1 shows that Phase I monitoring is very critical in process control efforts as it is designed to construct the monitoring system based on historical data. In practice, it is a common situation that the historical data consist of measurements from different sources or generated under different conditions. So there are potentially multiple change points in the data. Without identifying those change points, miss-leading results will be produced in Phase II monitoring. On the other hand, Phase I monitoring is a very challenging problem and not well studied in many applications (most monitoring work considers Phase II monitoring).

1.2 Research Problems

Quality monitoring is becoming an increasingly serious concern in manufacturing and service industries such as health care. In this research we have considered the Phase I quality monitoring problems in manufacturing and healthcare applications as described in the following.

1.2.1 Quality Monitoring in Low-E Glass Manufacturing

Low-E glass is a special kind of tempered glass where the heat is reflected back to the side where it was generated (Arasteh et al., 2004; Carmody et al., 1996; Frost et al., 1993). In summer when the heat is generated by the sun it is reflected back to the exteriors thus keeping the interiors of buildings cooler, while in winter when the heat is generated by the heaters it is reflected back to the interiors thus keeping the interiors warmer than the outside. The thermal emission of the glass is minimized by depositing various metals and metal oxides on the surface of the glass. The coating enhances the thermal properties of the glass by reflecting the infrared energy thus keeping the radiant heat on the same side of the glass where it was originated. As a result, the windows become more efficient because radiant heat originated from indoors in winter is reflected back inside, while infra-red heat radiation from the sun during summer is reflected keeping the inside cooler.

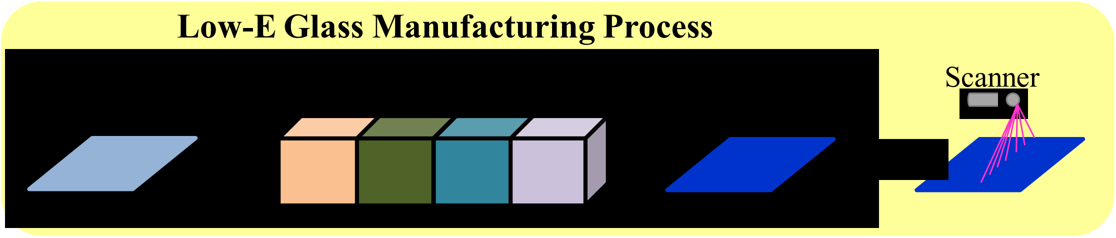


Figure 1-2 Low-E Glass Manufacturing Process

Figure 1.2 shows a schematic diagram of the Low-E glass manufacturing process. Glass ribbons as shown in the figure enter the coating chambers on one side.

Inside the coating chambers chemicals are deposited on the surface of the ribbon to enhance the optical properties of the glass. From the other end of the coating chambers the low-E glass exits. The finished product is scanned using scanners for quality control.

1.2.1.1 Quality Data in Low-E Glass Manufacturing: Optical Profiles

The quality data generated in Low-E glass manufacturing are optical profiles, as shown in Figure 1.3, which are one type of profile data. A profile, or a curve, represents the relationship of a response variable on an explanatory variable such as time and distance. In the optical profiles, the response is the optical property of the glass, i.e., the reflectance of light on the glass surface and the explanatory variable is the wavelength of light. Quality monitoring based on such data is called "Profile Monitoring". This is a new research topic in SPC field which has gained much popularity recently because quality profiles are becoming common in manufacturing processes Researchers in this field have done a lot of work on Phase II monitoring, while very little research has been done on Phase I analysis.

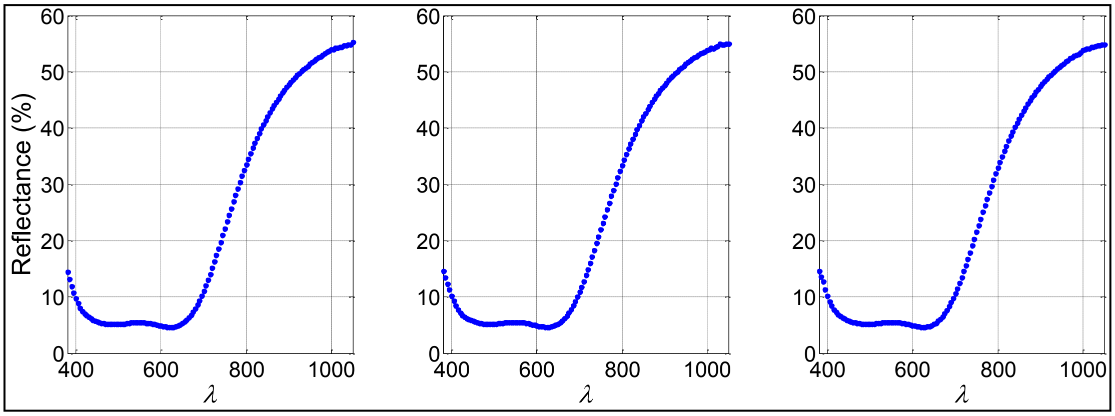


Figure 1-3 Optical Profiles

1.2.1.2 Problems in Phase I Monitoring of Profile Data: Non-normality

In the low-E glass manufacturing process, there are many chemical sub-processes that may generate various types of noises in the system. Those noises lead to deviation of the data from normality. In such a scenario, the existing statistical process monitoring techniques fail to provide a solution because most of them are based on the normality assumption of the data. Robust monitoring methods need to be developed to monitor the optical profile data in the low-E glass manufacturing process. Such methods have broad applicability as there are many other advanced manufacturing processes where normality is not satisfied.

This study is focused on the Phase I monitoring of optical profiles. There are three challenges in this problem. Firstly the data have high dimension. In general profile signals could contain as high as tens or hundreds of data points. For example, in each of the profile shown in Figure 1.3, there are 90 data points. How to deal with the high dimensionality of the data needs to be considered in the monitoring. The second challenge is to differentiate the out-of- control data from the incontrol data to identify the change points in the data. Especially, there might be multiple change points existing in the data. The true locations of those change points need to be identified with accuracy. Finally and most critically, since normality is violated in this case, a robust method that works on nonnormal data needs to be developed. This research proposes a robust strategy of Phase I analysis for optical profile monitoring. The proposed strategy aims to solve the abovementioned problems and provide a convenient process monitoring tool for practitioners in the low-E glass industry.

1.2.2 Risk-adjusted Readmission Monitoring in COPD Care

There are 3 components in this research: hospital readmission, chronic obstructive pulmonary diseases (COPD), and risk-adjustment. These three components are discussed in the following.

1.2.2.1 Hospital Readmission

The Patient Protect and Affordable Care Act (PPACA), also known as the Affordable Care Act (ACA) of 2010, became a law on March 2013. The main goal of this act is to improve the quality of health insurance. This health care reform has ensured that 30-day readmission be a metric to decide the quality of in-patient health care and a significant contributor to rising health care costs. A significant part of the healthcare act which focusses on reducing the costs related to readmissions poses a penalty to hospitals with high hospital readmissions.

However, as any chronic disease such as COPD advances, the condition of the patient becomes more severe. The patient may have more frequent exacerbations and hence more admissions to the hospitals. These factors can provide an estimate of how advanced stage COPD the patient has. Palliative care is usually started when a patient is on maximum medication yet his/her condition is getting worse. Palliative care means treatment to keep a patient as comfortable as possible in order to reduce the severity of the disease rather than to cure it. Most importantly, it helps the patient to bear. The goal of a palliative care is to focus on the planned care for the patient and his/her family. The idea is that in a hospital a multidisciplinary team of health care professional can anticipate any problems before they happen and help the patient with any medication and/or equipment needed.

1.2.2.2 COPD

COPD is a type of chronic disease which is a broad term for people with chronic bronchitis, emphysema, or both. Figure1.4 illustrates the classification of COPD.

- Chronic means tenacious and untiring.
- Bronchitis is the infection of the bronchi which are the airways of the lungs.

- Emphysema is injury to the smaller airways and alveoli (air pouches and bags) of the lungs.
- Pulmonary means anything affecting the lungs.

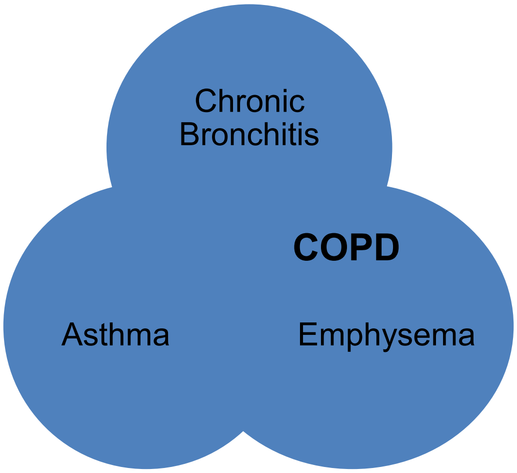


Figure 1-4 Classification of COPD

Patients with COPD have the airflow to the lungs restricted (obstructed). Symptoms in patients with COPD include cough and breathlessness. Chest infections are more common if a patient has COPD. A sudden worsening of symptoms when a patient has infection is called flare-up or exacerbation. Sputum generally turns yellow or green during a chest infection. Viruses that cannot be killed by antibiotics cause chest infections in COPD patients.

1.2.2.3 Risk Adjustment

Unlike products in industrial processes which are mostly homogenous, patients are heterogeneous because they come from different backgrounds, can have different conditions and severity that can add to the baseline risk. Thus, in monitoring the quality of care, we cannot only consider the patient outcomes such as the readmission of each patient, but also need to consider the patient risk factors, called “risk adjustment”. Risk-adjusted monitoring in health care has become a heated topic in the SPC field, and many risk-adjusted control charts have been developed which are extensions of their non-risk-adjusted counterparts in industrial applications.

1.2.2.4 Problems in Phase I Monitoring of COPD Readmission Data

The existing quality monitoring work in healthcare is limited in the following aspects:

First, they mostly focus on patient mortality in surgical/intensive unit care and no work has been done on the monitoring of patient readmission in chronic disease applications like COPD care. A special challenge in this new application area lies in the correlation in the readmission outcomes. Since a disease like COPD cannot be cured completely, one patient may be readmitted many times to the hospital. Consequently, the data may contain multiple observations on readmission from the same patients, which are intrinsically correlated. Such correlation needs to be considered in the monitoring. Second, the existing work only considers one covariate in the monitoring. Typically, the effects of all risk factors are summarized into one risk score and a simple logistic regression model is built to describe the dependency of the binary readmission outcomes on the risk scores. Then change detection is conducted based on this model. Obviously, this simple method cannot adequately model the effects of risk factors on readmission and may cause errors in monitoring.

Finally, most existing work on risk-adjusted monitoring focuses on Phase II analysis, and there is little work on Phase I analysis.

1.2.2.5 Summary of our Study and Contributions

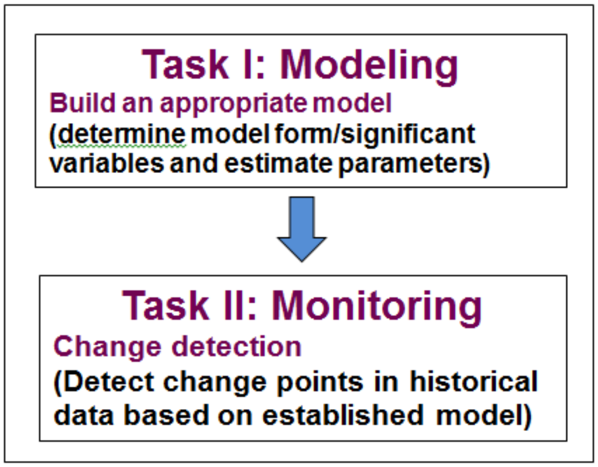

Task I: Modeling
Build an appropriate model
(determine model form/significant variables and estimate parameters)
Task II: Monitoring
Change detection
(Detect change points in historical data based on established model)

Figure 1-5 Proposed Approach

This research proposes a systematic approach for Phase I monitoring of patient readmission. It consists of two tasks as shown in Figure 1.5. In the modeling task, we build an appropriate statistical model for the data considering correlations in the readmission outcomes. Model/variable selection is done to determine the right form of model and significant risk factors to use in the modeling. Once the best model is determined, the parameters of the model are estimated. In the monitoring task, change detection based on the established model in the first task is examined and a convenient method for Phase I monitoring is developed.

This work contributes to the literature by solving the three issues mentioned in section 1.2.2.4: it is the first effort to consider monitoring of patient readmission in chronic disease care; it considers the effects of a large number of risk factors by proposing a systematic procedure for model building; and it proposes a method for Phase I risk-adjusted monitoring which can be applied in many areas in health care.

1.3 Outline of This Research

This report focuses on methodologies for addressing the two research problems described in section 1.2. Figure 1.6 shows the outline of this report.

Chapter 2 will present details of the proposed strategy of Phase I analysis for monitoring optical profiles in low-E glass manufacturing. A review of existing work on Phase I profile monitoring will be given and our proposed method will be explained. This is followed by the results and discussion of a simulation study and case study.

Chapter 3 is dedicated to the first task of the second research problem, i.e., risk-adjusted modeling of patient readmission in COPD care. Specifically, background introduction and literature review will be given first, and then details of the proposed approach will be presented. Results of the case study will be reported after that, where this approach is applied to a dataset from COPD patients. The established model will be used in the study of Task 2 in the future.

Chapter 4 will present the second task of the second research problem, i.e., risk-adjusted Phase I monitoring of patient readmission based on the model constructed in Chapter 3. A review of literature on risk-adjusted outcome monitoring in health care will be given and some issues in the future work will be discussed. The proposed method will be reported in my dissertation.

Chapter 5 will summarize studies and findings in this research and briefly describe directions of my future research.

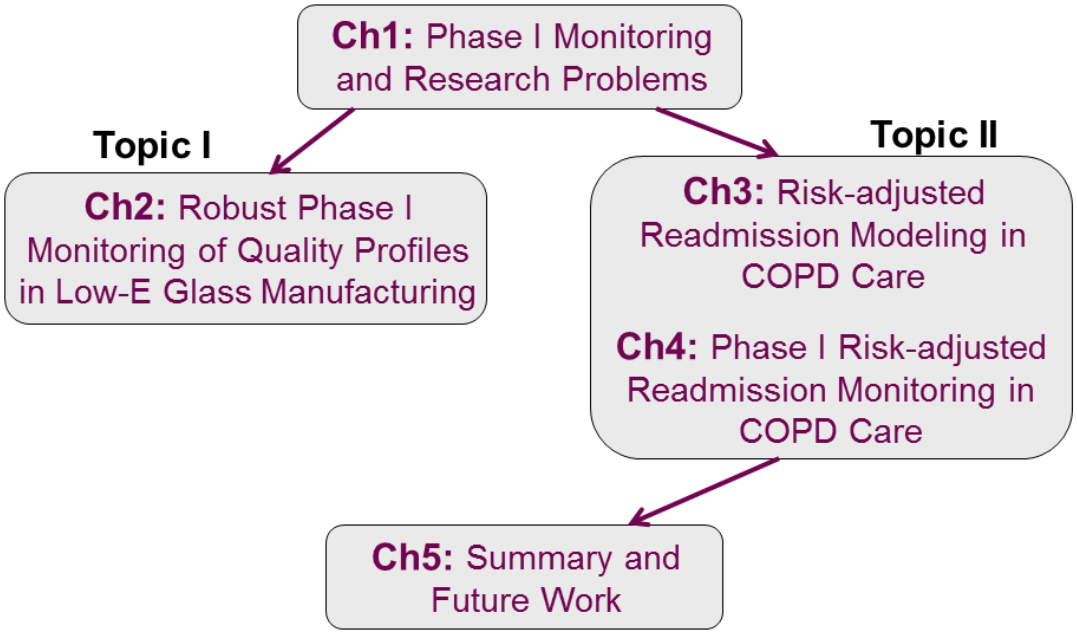

Ch1: Phase I Monitoring and Research Problems
Topic I
Topic II
Ch2: Robust Phase I Monitoring of Quality Profiles in Low-E Glass Manufacturing
Ch3: Risk-adjusted Readmission Modeling in COPD Care
Ch4: Phase I Risk-adjusted Readmission Monitoring in COPD Care
Ch5: Summary and Future Work

Figure 1-6 Outline of this Report

Chapter 2

Robust Phase I Monitoring of Profile Data with Application in Low-E Glass Manufacturing

2.1. Introduction

In some manufacturing processes, product quality is characterized by the relationship between a response variable and an explanatory variable which is called *profiles*. Monitoring of quality profiles has received much attention recently due to the increasing popularity of this type of quality data in practice (Woodall et al., 2004). Parametric and nonparametric methods have been developed for this purpose. This study focuses on parametric profile monitoring where the shape of the profiles can be characterized by a parametric model adequately.

The basic idea of parametric profile monitoring includes two steps: First, an appropriate statistical model is used to characterize the profiles. The choice of models depends on the characteristics of the profile data in the studied applications. Linear models (Kim et al, 2003), polynomial models (Kazemzadeh et al., 2009), splines (Walker et al., 2002), mixed-effect models (Jensen et al., 2008) and nonlinear models (Williams et al.,2007) have been used in existing studies. Second, the parameter estimates of the fitted model are monitored by using multivariate control chart techniques such as T^2 control chart and Multivariate EWMA control charts (Noorossana et al., 2011). Woodall (Woodall et al., 2007) and Noorossana et al. (2011) give excellent review of the state of art in this research area.

The majority of existing studies on profile monitoring focus on Phase II monitoring, while only a few efforts are made on Phase I analysis. Mahmoud and Woodall[13] develop an *F* test approach for Phase I monitoring of linear profiles. Mahmoud et al. (2007) propose a change point method based on likelihood ratio test for Phase I monitoring of linear profiles. Kazemzadeh et al. (2008) compare three approaches for

Phase I analysis of polynomial profiles, including the extension of the change point method, the extension of the *F* test approach, and a standard procedure based on T^2 test. It is found that the change point approach performs the best.

One limitation of the literature is that most studies rest on normality assumptions: the random errors in the profile models are typically assumed to be normally distributed, and the random effects in mixed-effect models are also bound by this assumption. However, this may not be the case in some manufacturing processes, such as the low-emittance (low-E) glass manufacturing process illustrated in Figure 2.1. The low-E glass is a type of energy-efficient glass products which is manufactured through physical or chemical coating processes, where solid materials, e.g., metal, metal oxide and metal nitride, are deposited on the surface of flat glass. The coating enhances the thermal/optical performance of the product so that they are able to reduce unwanted heat gain in summer and heat loss in winter (Arasteh et al., 2004, Frost et al., 1993 and Carmody et al., 1996). The quality of coating is measured by optical profiles of scanned locations on the glass surface. Figure 2.1 shows an example of a typical type of optical profiles, the reflectance profiles, which represent the percentage of light (r) that reflects from the glass surface over a range of wavelengths (λ). Due to the many chemical sub processes involved in low-E glass manufacturing, various random noises may be present in the production. As a result, the quality measurements may contain a considerable amount of extreme values and thus normality assumptions are not appropriate in modeling such profiles.

In fact, the effect of non-normality on the performance of profile monitoring has been investigated by a group of researchers such as Mahmoud and Woodall (2004), Williams et al. (2007), Vaghefi et al. (2009) and Noorossana et al. (2011). A general conclusion in these studies is that when normality is not satisfied, the conventional profile monitoring techniques may give misleading results. Though some techniques are found to be robust for certain types of deviations from normality (Noorossana et al., 2011) there lacks a generic method for profile monitoring in the presence of non-normality.

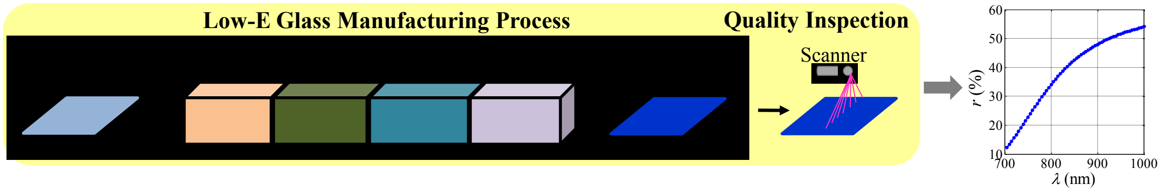


Figure 2-1 Low-E Glass Manufacturing Process and Example of Optical Profiles

This study aims to fill the gap in the literature by proposing a robust strategy for Phase I analysis of quality profiles. For non-normal data, nonparametric control charts are usually used to replace the conventional control charts based on normality. This idea is adopted in the proposed strategy. Moreover, to avoid issues with multivariate monitoring, independent component analysis (ICA) is used to transform multivariate coefficient estimates of the profile models into univariate independent data. In addition, we also study two methods to detect multiple change points as this is often the case in Phase I analysis. The properties of this strategy are demonstrated in a numerical study considering different scenarios of non-normality. In the case study, it is applied to optical profile data from low-E glass manufacturing as shown in Figure 2.1.

The remainder of the paper is organized as follows. Section 2.2 presents the problem formulation in Phase I analysis of profiles and the basic idea of the proposed strategy. Details of each component in the strategy are given in Section 2.3. Section 2.4 and 2.5 report the results of the numerical study and the case study respectively. Finally, Section 2.6 summarizes the findings in this work.

2.2 Problem Formulation

2.2.1 Profile Models and Phase I Monitoring

Without loss of generality, polynomial models which fit the optical profiles in Figure 2.1 will be used to illustrate the proposed method. Let *x* be the explanatory variable and *y* be the corresponding response. Suppose there are *m* profiles, each containing *n* sampling points, and the *x* values are fixed and constant among the profiles. Two types of polynomial models have been used in the literature: regular polynomial models (Kazamzadeh et al., 2008) and mixed-effect polynomial models (Amiri et al., 2009). Mathematical expressions of these models are given below:

Regular polynomial model

$$y_{ij} = \beta_p x_j^p + \ldots + \beta_h x_j^h + \ldots + \beta_0 + \varepsilon_{ij} \tag{1}$$

where i=1,...,m is the index of profiles, j=1,...,n is the index of sampling points, and h=0,...,p is the index of the exponent of polynomials. β_0,..., β_p are the fixed, unknown coefficients, and ε_{ij} is the random error which follows certain non-normal distribution with zero mean.

Mixed-effect polynomial model

$$\begin{gathered} y_{ij} = C_p x_j^p + \ldots + C_h x_j^h + \ldots + C_0 + \varepsilon_{ij} \\ C_h = \eta_h + \alpha_{hi} \end{gathered} \tag{2}$$

where each coefficient C_h consists of two parts: the *fixed effect*, η_h, which is fixed and unknown, and the random effect, α_{hi}, which varies from profile to profile. The random effects are assumed to follow non-normal distributions with zero mean. The mixed-effect models are preferred when the within-profile correlation is significant (Amiri et al., 2009 and Jensen et al., 2008) or when there is intrinsic variation in the shapes of profiles.

We take a change-point view in the Phase I analysis, that is, assume the historical data stream contains *w* change points, i.e.,

$$\begin{array}{ll} y_{ij} \sim M_1 & \text{for } 1 \le i \le K_1; \\ y_{ij} \sim M_2 & \text{for } K_1 + 1 \le i \le K_2; \\ \vdots & \vdots \\ y_{ij} \sim M_{w+1} & \text{for } K_w + 1 \le i \le m. \end{array} \qquad (3)$$

where $K_1,\dots, K_w$ are the change points, and $M_1,\dots, M_{w+1}$ are the polynomial models followed by the data between adjacent change points. Two types of changes may occur in the data: *location shift* which corresponds to the change in the coefficients in (1) or the fixed effects in (2), and *scale shift* which corresponds to the change in the scale of the random error in (1) or the scale of random effects/error in (2). The goal of Phase I monitoring is three-fold: determine whether any change occurs in the data, identify the change points as accurately as possible, and establish in-control parameters based on the change point estimates.

2.2.2 Basic Idea of the Proposed Strategy

Following the standard practice of parametric profile monitoring, change detection will be applied to the coefficient estimates of the profile models, i.e., βs in (1) or Cs in (2). When those estimates are not normally distributed, a natural idea is to use nonparametric control charts. As multiple coefficients typically exist, we can either use a multivariate nonparametric control chart on all the coefficients simultaneously, or use univariate nonparametric control charts on each coefficient separately. Since the estimates of coefficients are correlated with each other, the first solution appears to be more reasonable. It is also possible since multivariate nonparametric control charts are available in the literature, including the sign MEWMA chart proposed by Zou and Tsung (2011) and the rank-based MCUSUM chart and the nonparametric MCUSUM chart proposed by Qiu and Hawkins (2003). However, as calculating the statistics in these multivariate techniques involves matrix inversion operations, they may suffer instability issues in some cases. For example, according to our simulations, when the variation of the coefficient estimates is not balanced, that is, the variation of some coefficients is too large or too small, the statistics may not exist due to singularity of inverted matrices. In addition, the results from multivariate control charts do not have easy interpretation. In contrast, the second idea is free of instability issues and easy to implement and understand, given that the correlation between the estimates of coefficients can be eliminated in some way.

In fact, the second idea is followed in the study of Kazemzadeh et al (2009) for Phase II monitoring of polynomial profiles under normality assumptions, where orthogonal polynomial models are used for fitting the profile data. Since the coefficient estimates in orthogonal polynomial regression are independent, they can be monitored separately using univariate control charts. In this study, we propose a similar method to solve this problem in the context of non-normality, which uses independent component analysis (ICA) to transform the multivariate coefficient estimates into univariate independent components (ICs), and then applies univariate nonparametric control charts to each IC. This method is generic in that it can be applied to different forms of polynomial models and other models. Moreover, as will be explained in Section 2.3.2, the use of ICA will bring special benefits in change point detection.

Figure 2.2 shows the components of the proposed strategy for Phase I monitoring of profile data. First, polynomial models are fitted for the data to obtain the estimates of coefficients. Second, ICA is applied to the multivariate coefficient estimates. The selected ICs are then monitored using univariate nonparametric control charts to detect location/scale shifts. Considering that multiple change points may exist, once a change point is detected, the data stream will be segmented at the change point and the detection will continue on the uninspected data. Details of each component will be given in Section 2.3.

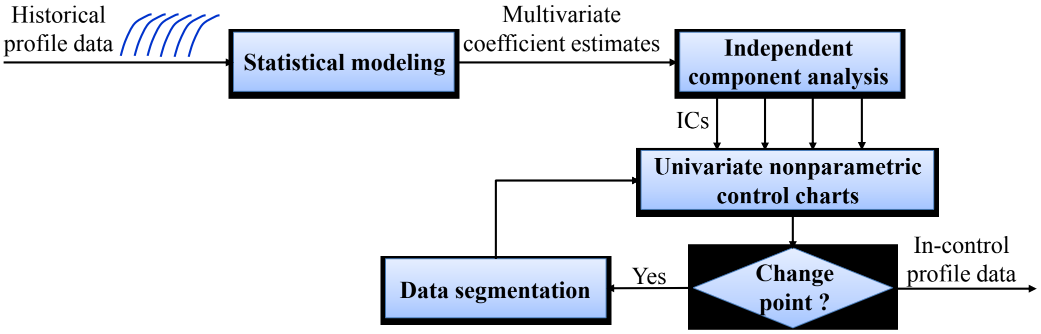


Figure 2-2 2Basic Idea of Proposed Strategy for Phase I Monitoring of Profile Data

2.3 The Proposed Strategy for Phase I Monitoring of Profile Data

2.3.1 Statistical Modeling

The first step in the Phase I analysis is to fit a polynomial model for each profile in the historical data stream. The degree of polynomials, *p*, can be determined through preliminary analysis comparing the residuals under different choices of *p*. Since non-normality is assumed, ordinary least squares method can be used to fit the models. If the underlying model is a regular polynomial model in (1), the estimates of coefficients are

$$\begin{bmatrix} \hat{\beta}_{p1} & \cdots & \hat{\beta}_{01} \\ \vdots & & \vdots \\ \hat{\beta}_{pm} & \cdots & \hat{\beta}_{0m} \end{bmatrix}$$

where each column represents the estimates of one coefficient from the *m* profiles. For mixed-effect models in (2), the obtained matrix represents the estimates of *C*s. The coefficient estimates will be used in the following analyses.

2.3.2 Independent Component Analysis

ICA is a data projection technique which transforms original multivariate data into univariate independent components through linear transformation (Hyvarinen et al., 2001). Similar to another popular projection tool, the principle component analysis (PCA), ICA is often used for dimension reduction purposes in the literature as a number of significant ICs can be selected to represent the original data. For example, it is used in monitoring complex nonlinear profiles to reduce the dimension of data points contained in each profile (Ding et al., 2006). Ding et al. point out another advantageous aspect of ICA: unlike PCA which projects data onto a lower subspace that preserves the majority of the variability in the original data, ICA projects data to a subspace where the distinction of any existing structures in the data will be maximized in the resulting ICs. So the objective of ICA aligns well with the objective of Phase I analysis, i.e., separating data following different structures.

With the abovementioned properties, ICA is appropriate in our study to transform the multivariate coefficient estimates into univariate independent components, so that univariate nonparametric control charts can be applied to detect changes. Some algorithms of ICA are available in commercial software such as Matlab and R. The *fastICA* function in Matlab is used in this study. It is worth mentioning that as the degree of polynomials is typically not high, the advantage of ICA in data reduction is not a key concern here. Results of the numerical study show clearly its role in manifesting the changes in the data, as given in Section 2.4.

2.3.3 Univariate Nonparametric Control Chart

Various univariate nonparametric methods have been developed for monitoring non-normal data, including the bootstrap control chart by Jones and Woodall (1998), and the rank-based tests by Gordon and Pollak (1994), and Hackl and Ledolter (1991). Here we choose the control chart proposed by Hawkins and Deng (2010) for detecting location shifts and the one proposed by Ross et al. (2011) for detecting scale shifts. These two techniques are chosen because they do not require prior knowledge of in-control parameters and easy to implement. Moreover, they can also be applied for Phase II monitoring of large-volume data streams which exist in many manufacturing processes such as the low-E glass manufacturing. Note that these techniques are designed to detect a single change point in the data; detection of multiple change points is realized through data segmentation which will be described in Section 2.3.4. The basics of the two techniques are provided as follows.

Assume $Z_1,\ldots, Z_m$ are independent non-normal random variables with distribution

$$\begin{aligned} &Z_i \sim F_1 \qquad \text{for } 1 \le i \le K; \\ &Z_i \sim F_2 \qquad \text{for } K+1 \le i \le m. \end{aligned}$$

where K is the change point between the two different distributions F_1 and F_2. The focus here is to determine whether a change exists and if so, estimate the change point K.

The control chart of Hawkins and Deng (2010) to detect location shifts is based on the Mann-Whitney two-sample test. Let

$$D_{ij} = \operatorname{sgn}(Z_i - Z_j) = \begin{cases} 1 & \text{if } Z_i > Z_j \\ 0 & \text{if } Z_i = Z_j \\ -1 & \text{if } Z_i < Z_j \end{cases}$$

where $1 \le i, j \le m$. The Mann-Whitney statistic is defined based on the D_{ij},

$$U_{k,m} = \sum_{i=1}^{k} \sum_{j=k+1}^{m} D_{ij}$$

for $1 \le k \le m-1$. The standardized version of this statistic is

$$U'_{k,m} = \frac{U_{k,m}}{\sqrt{k(m-k)(m+1)/3}}$$

which follows a standard normal distribution asymptotically. Note that this statistic holds for each possible value of k. A natural estimate of the change point is the value of k that gives the largest $U'_{k,m}$. In the control chart, the following statistic is used

$$\begin{aligned} U'_{\max,m} &= \max_{1\le k\le m-1} \left|U'_{k,m}\right| \\ \hat{K} &= \arg\max_{1\le k\le m-1} \left|U'_{k,m}\right| \end{aligned} \tag{4}$$

The control limit needs to be found through simulations under a specified in-control average run length (ARL$_{\text{in-control}}$). It is required that m≥15.

The control chart of Ross et al. (2011) to detect scale shifts is based on the Mood test. The Mood statistic is

$$M_{k,m} = \sum_{i=1}^{k} \left(R_i - \frac{m+1}{2} \right)^2$$

where R_i is the rank of Z_i among $\{Z_1,\ldots,Z_k\}$. The standardized version is

$$M'_{k,m} = \frac{\left|M_{k,m} - k(m^2-1)/12\right|}{\sqrt{k(m-k)(m+1)(m^2-4)/180}}$$

The statistic of the control chart takes a similar form as the Mann-Whitney statistic in (4),

$$\begin{aligned} M'_{\max,m} &= \max_{1\le k\le m-1} \left|M'_{k,m}\right| \\ \hat{K} &= \arg\max_{1\le k\le m-1} \left|M'_{k,m}\right| \end{aligned} \tag{5}$$

The control limit also needs to be obtained through simulations. Fortunately, Ross et al. (2011) provide polynomial approximations of the control limit under a group of in-control average run lengths. It is required that m≥20.

2.3.4 Data Segmentation for Multiple Change Point Detection

To identify multiple change points that may exist in the data, the two control charts in Section 2.3.3 need to be used repeatedly. There are two ways to do this as illustrated in Figure 2.3:

Binary segmentation (BS): Change detection is first conducted on all the data. Whenever a change is detected, the data stream is split into two segments at the estimated change point. Then change detection is conducted on each segment separately.

Sequential segmentation (SS): Change detection is conducted sequentially starting from the segment with minimum required number of data points. If no change is detected, a new data point will be added to the segment and the detection continues; when a change is detected, the segment by the estimated change point will be discarded and change detection is applied to the subsequent data.

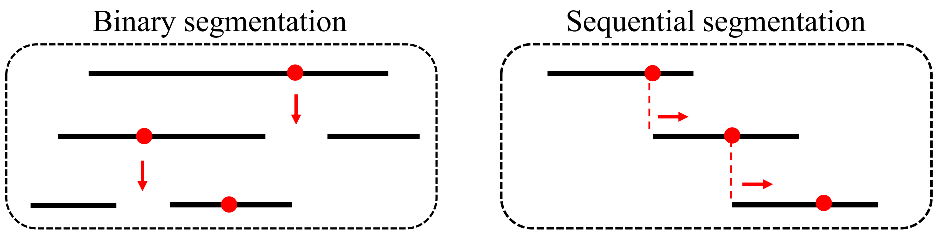


Figure 2-3 Two Ways for Data Segmentation in Multiple Change Point Detection

Each of the two methods has been used in existing studies for detecting multiple change points (Ross et al., 2011 and Kazemzadeh et al., 2008), but no study has been done to evaluate and compare their performance. In general, they both have pros and cons: the BS method works on a whole segment to detect changes, while the SS method adds new data point one by one. So the sample size in the BS method is likely to be larger than in the SS method, and thus the BS method tends to be more accurate in identifying the change points; on the other hand, the segment used in the BS method may contain multiple change points, while that used in the SS method is more likely to contain one single change point due to its sequential nature. So the assumption of single change point holds better for the SS method, and thus it is supposed to be more accurate. Simulation results on the performance of these two methods will be given in Section 2.4.2.

2.4 Numerical Study

Simulations are done to address the following concerns:

1. Performance of the two data segmentation methods described in Section 3.4 in multiple change point detection, and

2. Properties of the proposed strategy for Phase I monitoring of profile data.

For the first concern, univariate non-normal data streams containing two change points are simulated under different parameter scenario, and the two data segmentation methods are applied to each stream. Their performance in identifying the true change points is evaluated and compared. For the second concern, profile data with non-normal errors are simulated under different parameter scenarios, and the proposed Phase I analysis is applied. Characteristics of the proposed strategy will be summarized. In this section, we will first describe how data are generated in the simulations, and then report the results of the above studies.

2.4.1 Data Generation

Univariate data following non-normal distributions need to be simulated in this study. To be flexible, we use two large classes of non-normal distributions, the skew-normal distribution (Azzalini et al., 1985) and the skew-*t* distribution (Azzalini et al., 2008), which represent general cases of skewed and/or heavy-tailed distributions. For a random variable *Z* following the skew-normal distribution $SN(\mu, \sigma^2, \lambda)$ with location parameter μ, scale parameter σ^2 and skewness parameter λ, its density function has the following form

$$f(Z \mid \mu, \sigma^2, \lambda) = 2N\left(z \mid \mu, \sigma^2\right) \cdot \Phi\left(\lambda \frac{z-\mu}{\sigma}\right)$$

where *N*(*z*|*μ*, *σ*^{2}) is the density of normal distribution with mean *μ* and variance *σ*^{2}, and Φ is the cumulative distribution of the standard normal distribution. One issue with this parameterization is that it does not control the mean of *Z* directly so that the zero-mean assumption of the random errors/effects in model (1)-(2) cannot be implemented easily. To solve this problem, we adopt an alternative parameterization in the simulations

$$Z \sim SN(\omega, \tau^2, \lambda)$$
$$f(Z \mid \omega, \tau^2, \lambda) = 2N\left(z \mid \mu, \sigma^2\right) \cdot \Phi\left(\lambda \frac{z-\mu}{\sigma}\right)$$
$$\mu = \omega - \sqrt{\frac{2}{\pi}} \cdot \frac{1}{\sqrt{\frac{1+\lambda^2}{\lambda^2} - \frac{2}{\pi}}} \cdot \tau, \quad \sigma^2 = \frac{\tau^2}{1 - \frac{2}{\pi} \cdot \frac{\lambda^2}{1+\lambda^2}}$$

where *ω* and *τ*^{2} are the mean and variance of *Z*. Similarly, the skew-*t* distribution can be represented by

$$Z \sim ST(\omega, \tau^2, \lambda, v)$$
$$f(Z \mid \omega, \tau^2, \lambda, v) = \frac{2}{\sigma} t\left(z \mid \mu, \sigma^2, v\right) \cdot T\left(\lambda \frac{z-\mu}{\sigma} \sqrt{(v+1)/(v + \left(\frac{z-\mu}{\sigma}\right)^2)} \middle| v+1\right)$$

where *v* is the degree of freedom, and *μ* and *σ*^{2} can be obtained using the same formulas as in the skew-normal distribution.

Using the skew-normal and the skew-*t* distribution, we can simulate different situations of non-normality by manipulating their parameters. Sampling from these distributions can be done using Markov chain Monte Carlo (MCMC) algorithms (Robert et al., 2004). In our study, we use the slice sampler (Neal, 2003) through the *slicesample* function in Matlab to generate samples following the two distributions. Figure 2.4 shows the empirical distributions of examples of the simulated data, where *ω*=0, *τ*^{2}=1 and 100000 samples are generated in each case.

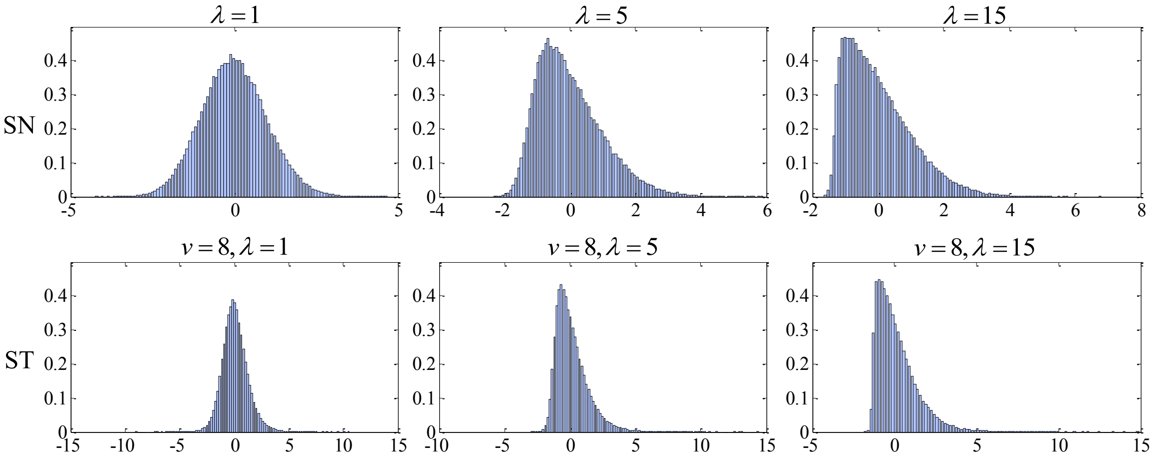
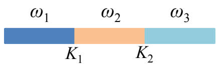
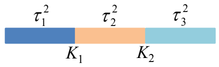


Figure 2-4 Normalized Histograms of Simulated Data from Skew-Normal and Skew-*t* Distribution

2.4.2 Performance of Data Segmentation Methods

In this study, data streams following skew-normal distribution (*λ*=6) and skew-*t* distribution (*λ*=6, *v*=6) are simulated. To obtain insight on the two data segmentation methods in multiple change point detection, a simple scenario is considered in which each data stream contains two equally-spaced change points (i.e., K_1=100, K_2=200, *m*=300) or in other words, three segments with equal length (100). The changes are either location or scale shifts. When location shifts occur, the scale parameters of the three segments take the same value (τ^2=1), while their location parameters ω_1, ω_2, and ω_3 are different. Similarly, when scale shifts occur, the location parameters of the three segments are the same (*ω* =0), while their scale parameters τ_1^2 , τ_2^2 and τ_3^2 take different values. 4 cases are simulated under each type of shifts, which lead to a total of 8 cases. Table 2.1 summarizes the parameter settings and interpretations in these cases. Figure 2.5 shows an example of data streams generated in each case, where the solid line in each plot indicates the true value of the location parameter.

Table 2-1 Parameter Settings in Evaluating Performance of Data Segmentation Methods

Case	 **Location shift**			 **Scale shift**			**Interpretation**
	ω_1	ω_2	ω_3	τ_1^2	τ_2^2	τ_3^2	
I	0	1	0	1	2.5	1	a small change, then back to in-control
II	0	2	0	1	4	1	a large change, then back to in-control
III	0	2	1	1	4	2.5	a large change, followed by a small change
IV	0	1	2	1	2.5	4	a small change, followed by a large change

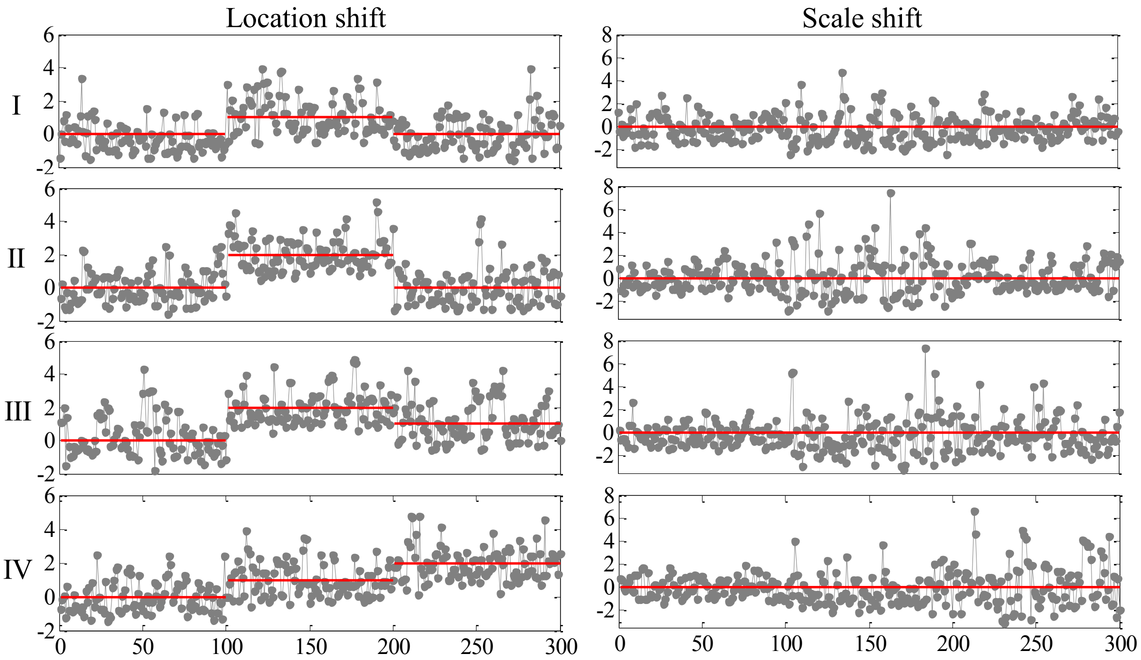


Figure 2-5 Examples of Data Streams Generated under each case listed in Table 2.1

Under each case listed in Table 2.1, 10000 data streams are simulated. The BS and the SS method are applied to each of the streams. In using the control charts in (4) and (5), a control limit with ARL$_{\text{in-control}}$=2000 is applied. The performance of the two methods in each case is evaluated using the following measures:

$$R_{FA} = probability\ that\ more\ than\ 2\ change\ points\ are\ detected$$

$$R_{MIS} = probability\ that\ only\ 1\ change\ point\ or\ no\ change\ point\ is\ detected$$

$$R_{E1} = probability\ that\ the\ change\ point\ estimate\ is\ within\ 10\%\ interval\ of\ K_1$$

$$R_{E2} = probability\ that\ the\ change\ point\ estimate\ is\ within\ 10\%\ interval\ of\ K_2$$

Here the “10% interval” in the last two measures means the interval [90, 110] for K_1, and [190, 210] for K_2. The above performance measures essentially represent the false alarming rate, miss detection rate, and accuracy in estimating K_1 and K_2.

Table 2.2 gives the results on the performance measures in location shift detection. Figure 2.6 shows the corresponding distributions of change point estimates for skew-normal data. The change point estimates for skew-*t* data exhibit similar patterns. We find the following things from the results:

The performance of the BS and the SS method shows some common characteristics: According to results in Table 2.2, both methods have lower miss detection rate and more accurate change point estimates when the location difference between the two sides of the change point is higher. From the upper panel of Figure 2.6, we can see that the estimates of K_1 in Case II and III have a sharper distribution than in other cases, meaning that the estimation is more accurate. This is because the difference in the locations at the two sides of K_1 is larger in these two cases. For the estimation of K_2, Case II performs the best as the location difference at the two sides of the change point in this case is larger than in other cases.

The two methods are different in two aspects: (1) From Table 2.2, the BS method has much smaller false alarming rate and considerably larger miss detection rate than the SS method. This means that the BS method tends to miss some change points, while the SS method tends to detect some false change points. This is consistent to our intuitive understanding of these two methods given in Section 2.3.4: since the BS method works on a whole segment which contains more information, it is less likely to signal a false change point; but meanwhile it is more likely to miss some true change points as it can only pick one change point from the segment being inspected which may in fact contain multiple change points. In contrast, the SS method examines the data sequentially so that it is more likely to detect the true change points; but meanwhile it tends to generate more false alarms due to the limited information used especially at the beginning of each detection. (2) From Figure 2.6, we can see that the change point estimates from the two methods have similar distributions in general, with the mode of the SS method being slightly higher than the BS method. Overall we can say that they provide change point estimates of similar accuracy.

Comparing the skew-normal and skew-*t* data: The results of the two distributions show similar patterns, but in most cases the skew-*t* data have higher false alarm rate and miss detection rate, and less accurate change point estimates than the skew-normal data.

Table 2-2 Performance of the BS and the SS Method in Detecting Location Shifts

	Case	**Binary segmentation**				**Sequential segmentation**			
		R_{FA}	R_{MIS}	R_{E1}	R_{E2}	R_{FA}	R_{MIS}	R_{E1}	R_{E2}
SN	**I**	0.1049	0.0234	0.8081	0.8034	0.4722	0.0002	0.8711	0.8720
	II	0.1306	0	0.9841	0.9834	0.5075	0	0.9830	0.9969
	III	0.0724	0.0011	0.9797	0.8848	0.5068	0.0003	0.9849	0.8708
	IV	0.1655	0.0064	0.8281	0.8132	0.4481	0.0002	0.8680	0.8678
ST	**I**	0.1476	0.0938	0.6804	0.6820	0.6526	0.0025	0.8019	0.8075
	II	0.1885	0	0.9395	0.9403	0.6676	0	0.9477	0.9827
	III	0.1305	0.0106	0.9393	0.8026	0.6718	0.0005	0.9538	0.8030
	IV	0.1948	0.0452	0.7301	0.7219	0.6139	0.0040	0.8062	0.8020

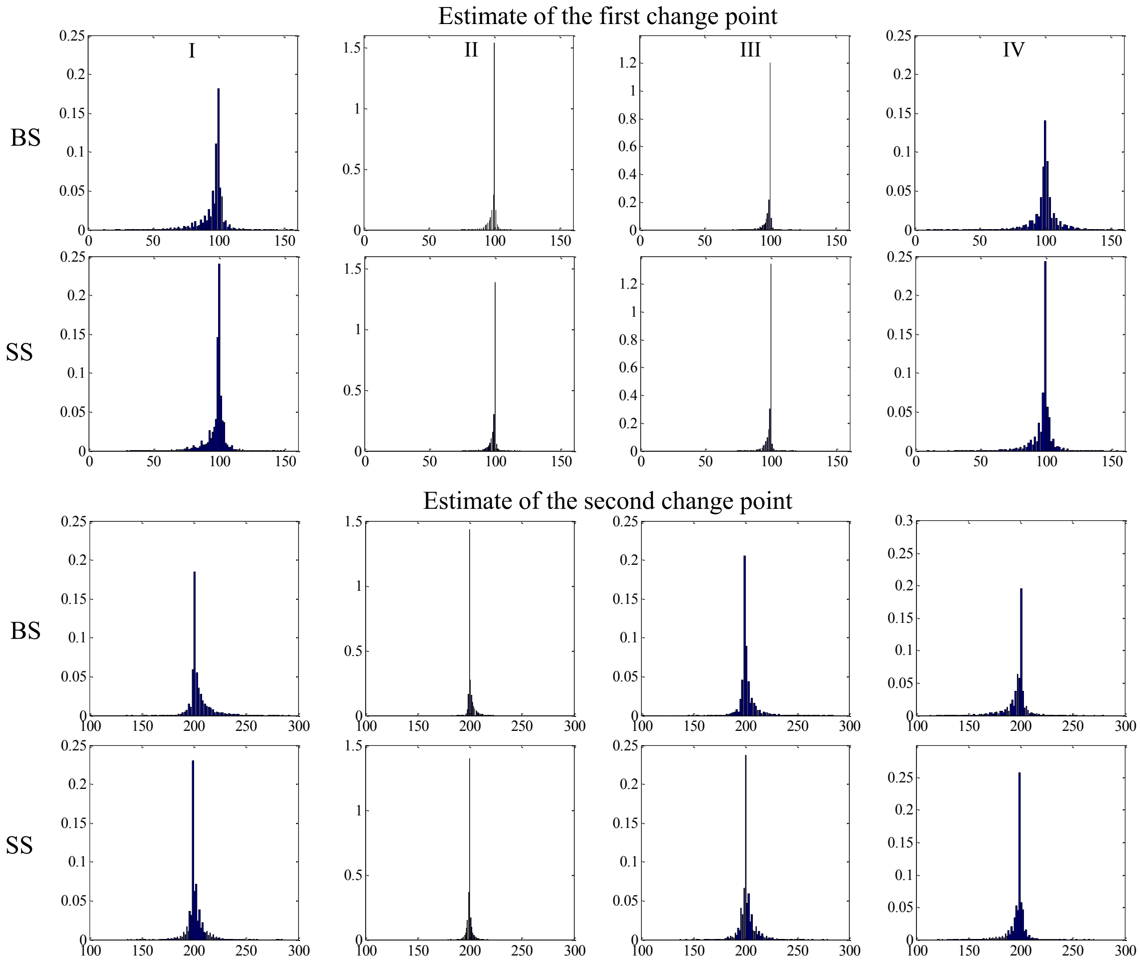


Figure 2-6 Normalized Histograms of Change Point Estimates Under Location Shifts

The results on the performance measures in detecting scale shifts are given in Table 2.3, and the corresponding distributions of change point estimates for the skew-normal data are shown in Figure 2.7. In general, the performance of the two methods shows similar patterns as in the cases of location shifts. Both methods perform the best in Case II where the difference at the two sides of the change points is larger than in other cases. The BS method has higher miss detection rate, while the SS method has higher false alarming rate. Both rates are larger than in the cases of location shifts.

Correspondingly, the distribution of change point estimates has larger variance. This is because scale shifts are, in general, more difficult to detect than location shifts.

Table 2-3 Performance of the BS and the SS Method in Detecting Scale Shifts

	Case	Binary segmentation				Sequential segmentation			
		R_{FA}	R_{MIS}	R_{E1}	R_{E2}	R_{FA}	R_{MIS}	R_{E1}	R_{E2}
SN	**I**	0.0702	0.5057	0.3470	0.3460	0.6097	0.0256	0.5994	0.5743
	II	0.1341	0.0757	0.7613	0.7618	0.6563	0.0007	0.8016	0.8137
	III	0.0411	0.5819	0.8383	0.1378	0.5213	0.2055	0.7737	0.2420
	IV	0.0280	0.7749	0.4958	0.1259	0.5918	0.1367	0.5895	0.2779
ST	**I**	0.0890	0.5385	0.3032	0.3072	0.6715	0.0324	0.5381	0.5319
	II	0.1670	0.1051	0.6943	0.6924	0.7195	0.0016	0.7672	0.7580
	III	0.0651	0.5775	0.7730	0.1312	0.6157	0.1547	0.7420	0.2575
	IV	0.0419	0.7679	0.4404	0.1296	0.6522	0.1223	0.5469	0.2855

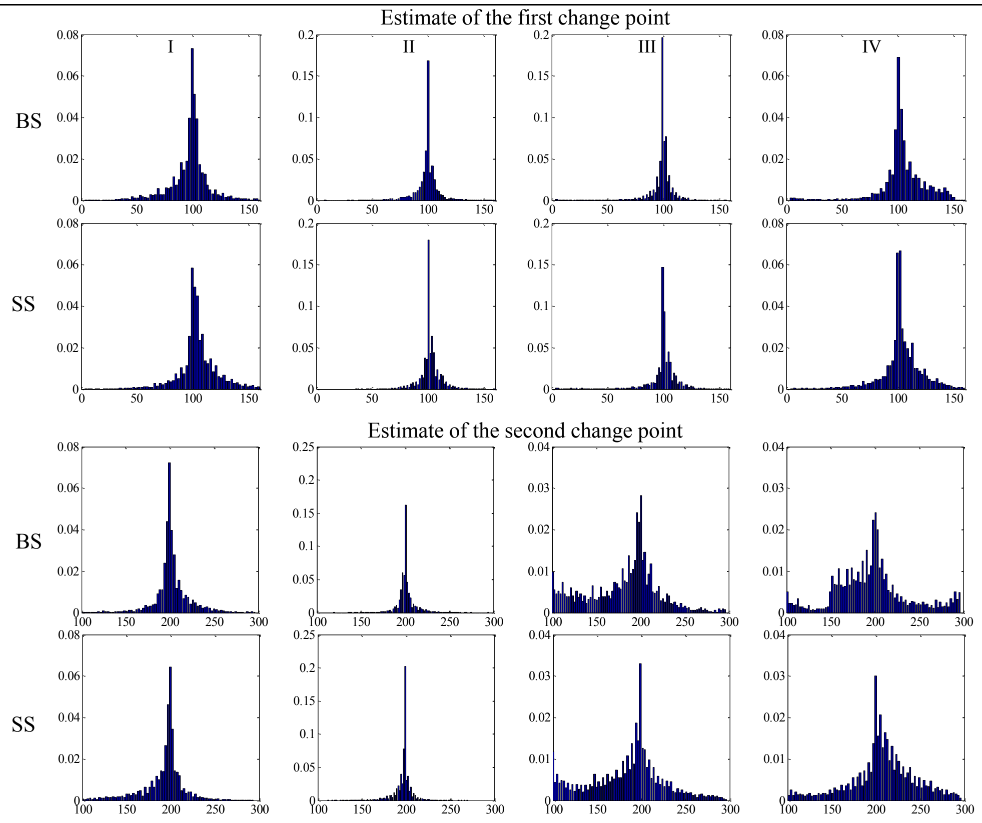


Figure 2-7 Normalized histograms of change point estimates under scale shifts

2.4.3 Properties of the Proposed Strategy for Phase I Monitoring

In this study, we simulate streams of profile data following the regular polynomial model in (1) or the mixed-effect model in (2) with degree-2 polynomials and apply the proposed strategy for Phase I monitoring on each stream. Like the simulated data in Section 2.4.2, each stream contains three segments, and each segment contains 100 profiles following a different model. The random errors/random effects in the models are generated from skew-normal distributions. The in-control models are

Regular polynomial model: $y = \beta_2 x^2 + \beta_1 x + \beta_0 + \varepsilon$

$$\beta_2 = \beta_1 = \beta_0 = 2, \;\; \varepsilon \sim SN(\omega = 0, \tau^2 = 1, \lambda = 6)$$

Mixed-effect polynomial model: $y_i = \eta_2 x^2 + \eta_1 x + \eta_0 + \alpha_{2,i} x^2 + \alpha_{1,i} x + \alpha_{0,i} + \varepsilon$

$$\begin{aligned}
&\eta_2 = \eta_1 = \eta_0 = 2 \\
&\alpha_{2,i} \sim SN(\omega = 0, \tau_2^2 = 1, \lambda = 6) \\
&\alpha_{1,i} \sim SN(\omega = 0, \tau_1^2 = 1, \lambda = 6) \\
&\alpha_{0,i} \sim SN(\omega = 0, \tau_0^2 = 1, \lambda = 6) \\
&\varepsilon \sim SN(\omega = 0, \tau_\varepsilon^2 = 1, \lambda = 6)
\end{aligned}$$

where the explanatory variable *x* takes values [0, 0.1, 0.2,…., 3.0]. To be convenient, the change structure in Case I and II in Table 2.1 is applied to each data stream, that is, the first and third segments follow the above in-control model, while the second segment follows a different model. 6 cases are simulated considering different settings of the parameters of the second segment, which are listed in Table 2.4. Under each case, profile streams are generated and the proposed Phase I analysis is applied to each stream. The results of one typical example under each case are shown in Figure 2.8. In each plot of the figure, the left column displays the estimates of coefficients, while the right column displays the selected independent components. The estimated change points are marked in the plots of ICs.

Table 2-4 Parameter settings of the second segment in the simulated profile data

Case	Model	Parameters	Interpretation
1	Regular	$\beta_2 = 2.5$	Small location shift in quadratic coefficient
2	Regular	$\beta_1 = 3$	Mild location shift in linear coefficient
3	Regular	$\beta_1 = 2.5$, $\beta_0 = 2.5$	Small location shift in both linear coefficient and intercept
4	Regular	$\tau^2 = 2.5$	Small scale shift
5	Regular	$\lambda = 15$	Mild shift in skewness
6	Mixed-effect	$\tau_1^2 = 8$	Large shift in random-effect variance

The results in Figure 2.8 can be summarized in the following aspects

1) The effect of ICA: We can see that the shifts manifest themselves more clearly in the ICs than in the coefficient estimates. This is particularly the case in Figure 2.8(c) where the data contain two small location shifts. Little evidence of the shifts can be found in the coefficient estimates, while the evidence is quite apparent in the ICs. This validates the intrinsic capacity of ICA in manifesting the structure in the data. Another observation is that the shifts tend to appear in the first ICs, which implies the potential of ICA for data reduction when a large number of coefficients exist.

2) Change point estimation: From Figure 2.8(a)-(c), it is seen that the change points of location shifts are estimated accurately. Not surprisingly, from Figure 2.8(d), we see that it is more difficult to estimate change points of scale shifts than location shifts. In Figure 2.8(f), due to the random effects of the coefficients in the mixed-effect model, Change point estimation: From Figure 2.8(a)-(c), it is seen that the change points of location shifts are estimated accurately. Not surprisingly, from Figure 2.8(d), we see that it is more difficult to estimate change points of scale shifts than location shifts. In Figure 2.8(f), due to the random effects of the coefficients in the mixed-effect model, estimation of the change points in scales becomes even more difficult. But according to our simulations not

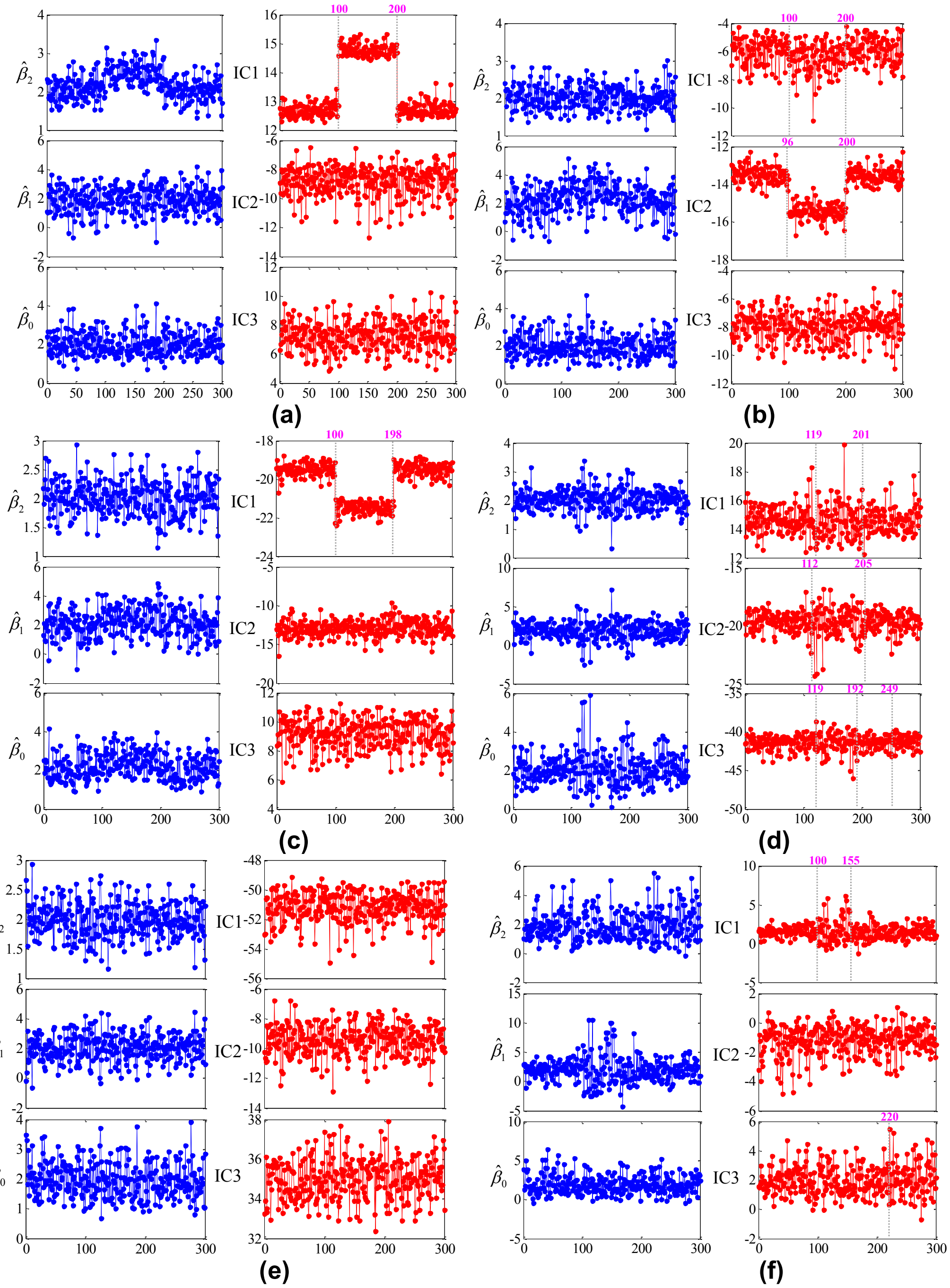


Figure 2-8 Example of Coefficient Estimates and Selected ICs under Each Case (a-Case-1, b-Case-2, c-Case-3, d-Case-4, e-Case-5, f-Case-6 listed in Table 2.4

shown here, the accuracy in the estimation gets improved when the magnitude of the shift is larger. Finally, as shown in Figure 2.8(e), the two nonparametric control charts cannot detect shifts in skewness, which is reasonable as they are designed for location/scale shifts

2.5 Case Study

In this study, the proposed Phase I analysis is applied to a set of optical profile data as shown in Figure 2.1. The data were from a large-scale low-E glass producer in the US. For confidentiality reasons, the name of the company and information of their products are not disclosed in this text. The data set consists of 314 optical profiles and each profile contains 30 data points corresponding to λ=[705nm, 710nm, …,850nm]. Before implementing the analysis, some preprocessing is done on the raw data. This includes the centering/scaling transformation of λ values, i.e., x=[λ–average(λ)]/150, which can improve the numerical properties of the fitting, and determining the appropriate degree of polynomials through fitting polynomial models to each profile and checking the residuals. As an example, Figure 2.9 shows the fitted models and resulting residuals for one profile. We can see that the residuals become very small and exhibit random patterns with equal variance when p=4. Therefore, we decide that the degree-4 polynomial model gives adequate fitting and will be used in the Phase I analysis First, coefficient estimates are obtained from each profile, which are shown in Figure 2.10. The estimates consist of a considerable amount of extreme values, a sign of non-normality. It appears that multiple change points may exist in the data, and an apparent one of which occurs during profiles #200~#250. Then ICA is applied to these estimates. Figure 2.11 shows the resulting ICs. The apparent shift can be seen in the first IC, and there is also evidence of shifts in other ICs.

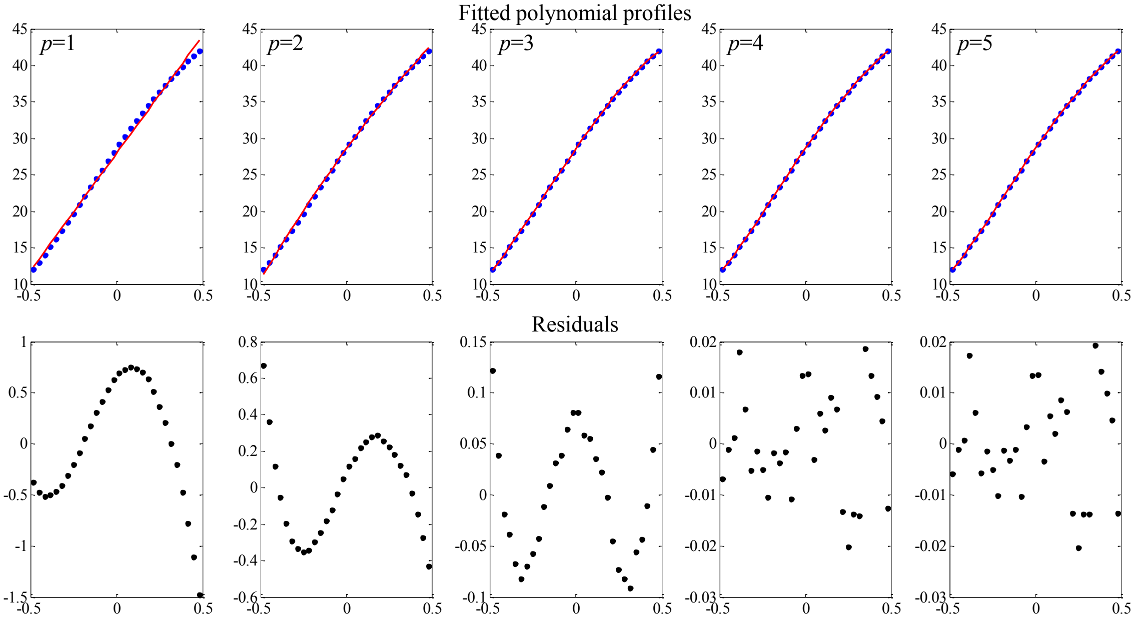


Figure 2-9 Example of fitted polynomial models and residuals

The BS and the SS method are applied to each IC. The estimates of change points are listed in Table 2.5. As expected, more change points are detected by the SS method, especially in detecting scale shifts. But the change point estimates from the two methods are very similar. For location shifts, multiple change points are detected including the apparent one (#232) in Figure 2.10. Fewer change points are detected for scale shifts. Particularly, only one change point is obtained by the BS method. Using the detected change points, the data are divided into multiple segments. Figure 2.12 and 2.13 show the segments based on the results of the SS method.

Table 2-5 Estimates of change points for each independent component

IC	**Location shift detection**		**Scale shift detection**	
	Binary seg.	Sequential seg.	Binary seg.	Sequential seg.
IC1	55, 100, 159, 232	55, 100, 159, 229	234	5,160, 232
IC2	89, 148, 232, 246	89, 148, 232, 246	N/A	36, 48,246
IC3	50, 140, 304	50, 140, 291	N/A	N/A
IC4	14, 100, 122, 162, 232, 248	14, 99, 122, 159, 232	N/A	118, 251
IC5	69, 128	13, 69, 128, 158, 247	N/A	159

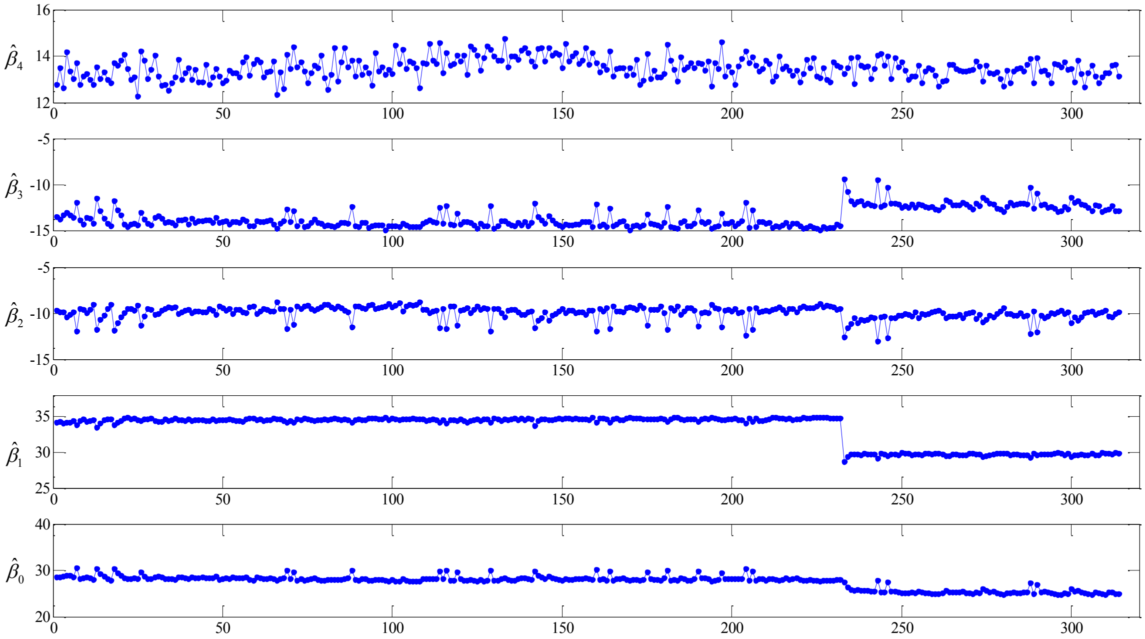


Figure 2-10 Estimates of Coefficients of Degree-4 Polynomial Models

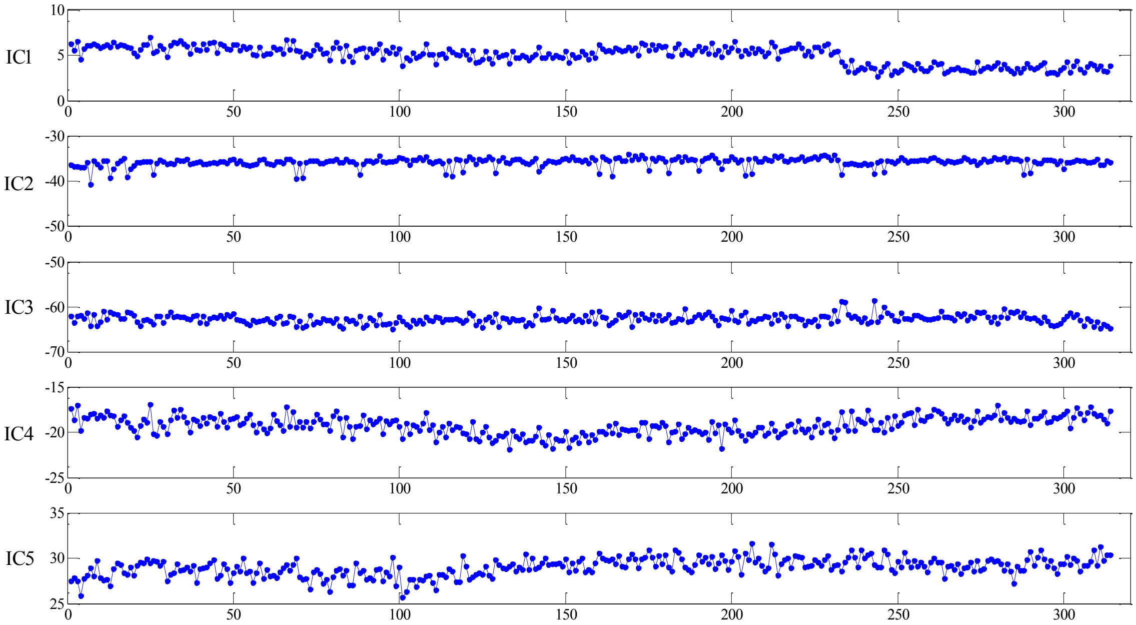


Figure 2-11 Independent Components Obtained from the Coefficient Estimates

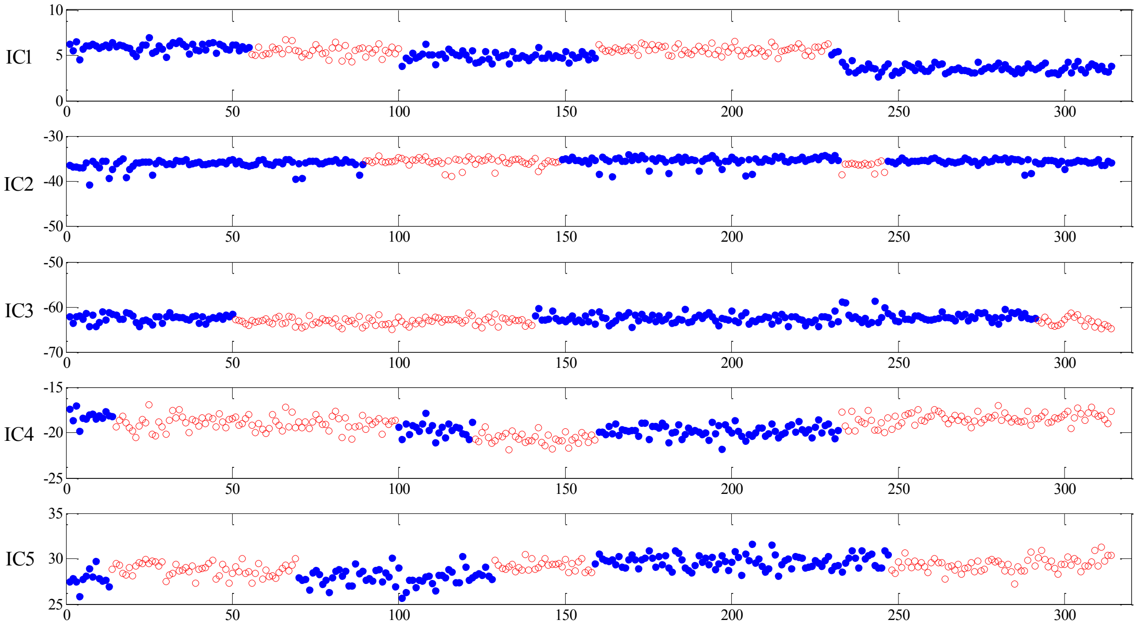


Figure 2-12 Estimates of Location Change Points using the Sequential Segmentation Method

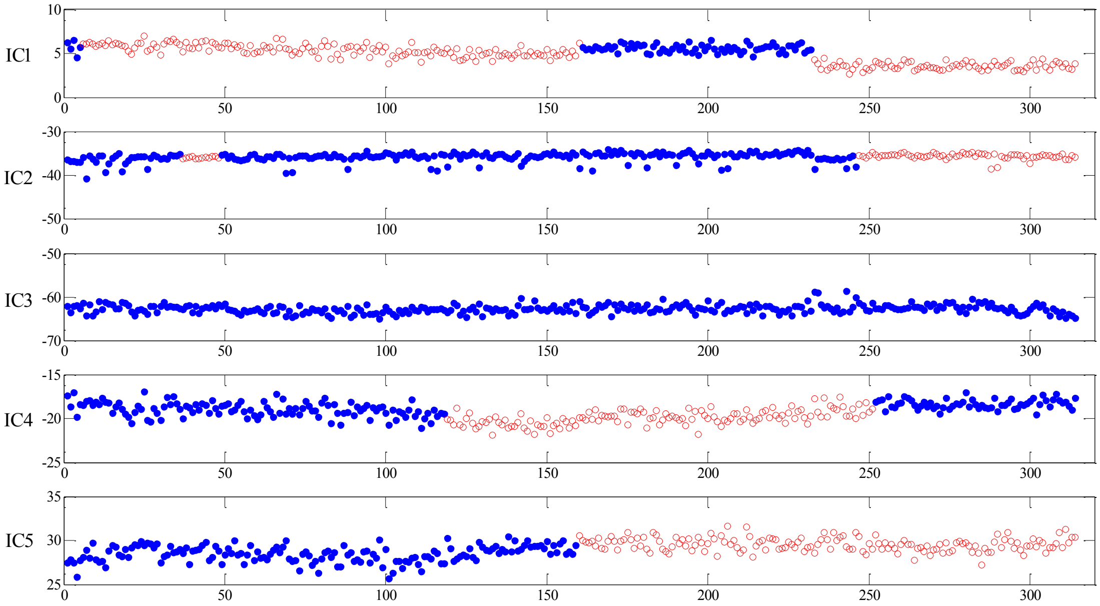


Figure 2-13 Estimates of Scale Change Points using the Sequential Segmentation Method

Based on the results of the two methods, two groups of profiles are identified which have constant location and scale. The two groups contain profiles #159~#232 and #246~#291, which are shown in Figure 2.14. They have apparently different shapes, indicating that the process underwent a location shift. Degree-4 polynomial models are fitted for the two groups separately. Figure 2.15 shows the quantile-quantile (QQ) plots of the coefficient estimates of the first group. It is clear that the distribution of the estimates is not normal, which justifies the use of non-parametric change detection techniques.

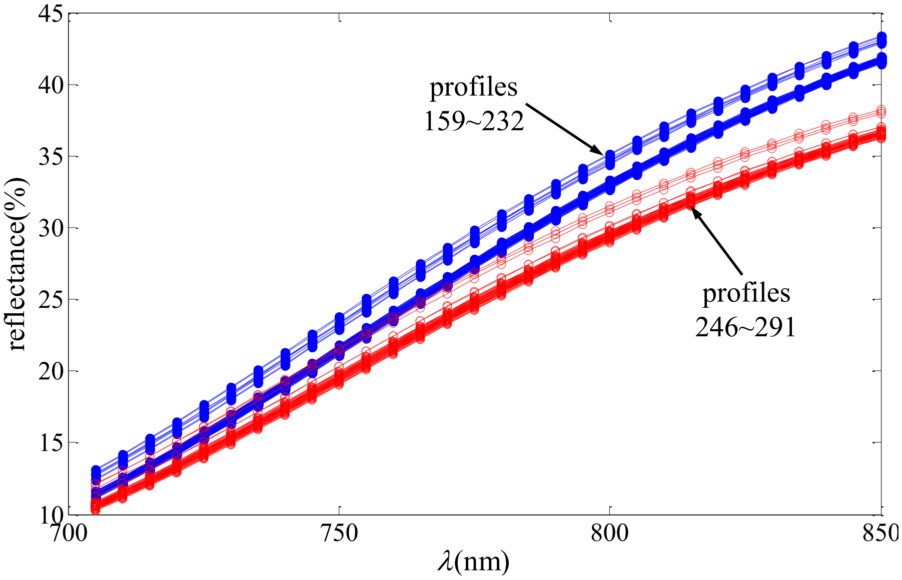


Figure 2-14 The Identified Two Groups of Profiles

We have consulted with the engineers on the findings in the Phase I analysis. After carefully reviewing the process history, they identified an abrupt change in the voltage/current of certain coating chambers which is likely to have caused the location shift between the two groups of profiles in Figure 2.14. It is believed that such changes occur when the production switches to a different type of glass products. They also captured a number of small drifts in some process variables such as the oxygen/nitrogen flow in certain chambers. These drifts are likely to be the reason for other change points shown in Figure 2.12 and 2.13. Such drifts are in general difficult to control due to all sorts of random factors in the coating process.

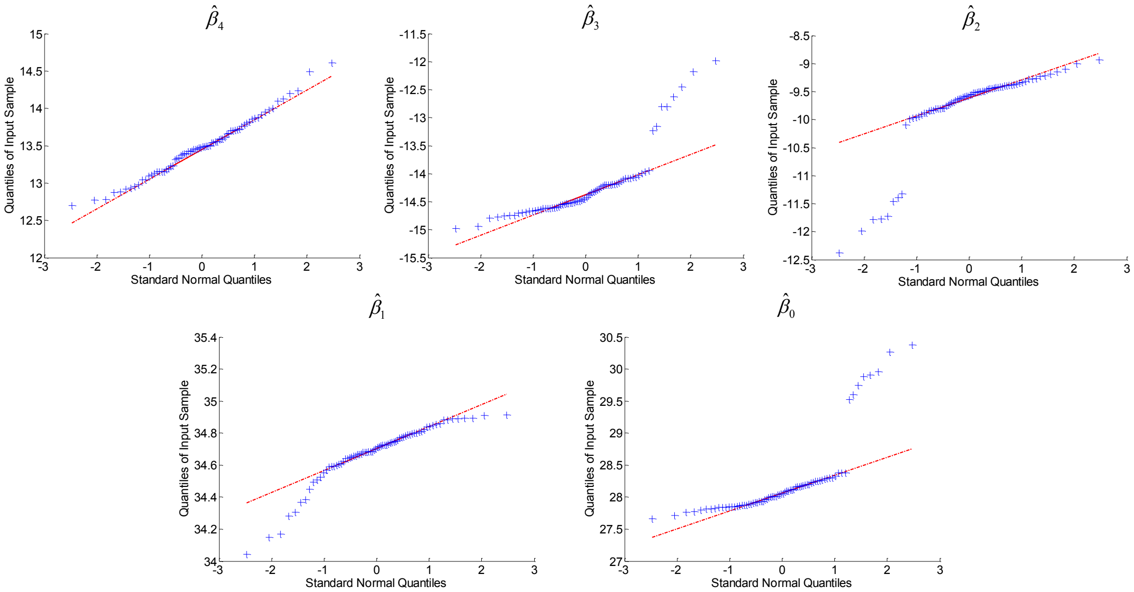


Figure 2-15 QQ-Plots of the Coefficient Estimates of Profiles #159~#232

2.6 Summary

This study proposes a strategy for Phase I monitoring of profile data under non-normality. The strategy contains three components: fitting appropriate models for profiles, independent component analysis on coefficient estimates, and change point detection on each independent component using nonparametric control charts. The performance of this strategy is studied through simulations on general classes of non-normal distributions. It is found that the use of ICA can reveal the structure in the data; between the two methods to detect multiple change points, binary segmentation has a lower false alarming rate, while sequential segmentation has a lower miss detection rate; the estimation of change points is more accurate when the difference in the location/scale at the two sides of the change point is larger. In the case study, the proposed strategy is applied to optical profiles from low-E glass manufacturing. A number of change points are detected, and two groups of profiles with constant location/scale are identified. Causes for the detected process shifts are also analyzed.

Chapter 3

Risk-Adjusted Modeling of Patient Readmission in COPD Care

3.1 Literature Review

There are two major components in this research, i.e., patient readmission and statistical models for binary readmission data. Many studies have been done on these topics. A brief review of the literature is given in this section.

3.1.1 Patient Readmission

Many studies have been done to identify risk factors and build prediction models for hospital readmission in various medical applications. In those studies readmission is typically measured either by binary indicators of whether a patient had readmission within a short period after discharge, e.g., 30 days, or by the interval between discharge and readmission.

Kariv et al. (2006) identified the risk factors for readmission after major abdominal surgery that may improve postoperative care and discharge plans. Stewart et al. (2000) identified risk factors for 30 day hospital readmission following Coronary Artery Bypass Grafting (CABG). Ferraris et al. (2001) investigated the factors associated with early hospital readmission after cardiac procedures. The idea is to develop strategies to minimize the problem. Kiran et al. (2004) determine the readmission rate and outcomes for patients undergoing intestinal operations. Variables that might predict readmission are evaluated.

There are also many studies on the readmission of COPD patients. Kansagara et al. (2011) summarize validated readmission risk prediction models, describe their performance and assess their suitability for clinical or administrative use. COPD readmission is perceived as an adverse effect in itself and suggested to be used as health service performance indicator for quality monitoring. COPD is the third leading cause of death in the United States and the only leading cause for which morbidity and mortality are rising. Over half of the patients who are hospitalized for acute exacerbations are readmitted at least once in the ensuring 6 months. Cao et al. (2006) ascertain rates of re-hospitalizations for AECOPD patients and evaluate factors associated with frequent readmissions for acute exacerbations. They find that frequent past readmission for AECOPD is associated with disease severity and psychosocial distress and increased use of vaccinations. Hospitalizations for such patient accounts for as much as 40% of the total direct cost of medical care of COPD in the nation.

Garcia et al. (2007) suggest that the Integrated Care (IC) intervention improved the COPD disease knowledge and treatment adherence suggesting that the factors such as education, coordination among levels of care, and improved accessibility, reduced hospital readmission in COPD after 1 year. Lau, Yam and Poo (2001) find out the factors associated with shorter time to first readmission after discharge from hospital after acute exacerbation. The factors considered are demographic and social data, comorbidities, treatment and first blood investigation after admission. Chen, Li and Johansen (2001) compare factors such as sex and age in hospital readmissions for COPD associated with overall and cardiac comorbid conditions. Gudmundsson et al. (2005) present a study to analyze the risk of re-hospitalization in patients with COPD disease and associated risk factors. They find that in patients with low health status, anxiety is an important risk factor for rehospitalization. Puhan et al. (2005) find evidence from their trials that respiratory rehabilitation is effective in COPD patients after acute exacerbation. Almagro et al. (2006) identify the risk factors for hospital readmission in COPD patients. In bivariate analysis the readmission is found to be associated with previous hospitalizations. In multivariate analysis the best predictor of readmission is found to be the combination of hospitalization for COPD in the previous year and $PaCO_2$ at discharge. Hasan et al.

(2004) identify predictors of early hospital readmission in a diverse patient population. They find that seven significant predictors of early readmission: insurance status, marital status, having a regular physician, Charlson comorbidity index, SF12 physical component score ≥ 1 admission within the last year, and current length of stay > 2 days. Bahadori and FitzGerald (2007) use systematic review to summarize the results from available studies to identify potential risk factors for hospital admission and/or readmission among patients experiencing COPD exacerbations. Ng et al. (2007) evaluate the impact of comorbid depression on mortality, hospital readmission, smoking behavior, respiratory symptom burden, and physical and social functioning in patients with COPD. Smith et al. (2000) determine clinical and patient-centered factors predicting non-elective hospital readmission. They find that the risk of readmission increases if the patient has more hospitalizations and emergency room visits in the prior 6 months, higher blood urea nitrogen, lower mental health function, a diagnosis of OCPD and increased satisfaction with access to emergency care assessed on the index hospitalizations. Garcia-Aymerich et al. (2003) find the factors causing exacerbations in COPD. Their final multivariate model shows the risk factors such as admissions for COPD in the year before recruitment, FEV1, percentage predicted, oxygen tension, higher levels of usual physical activity and taking anticholinergic drugs. Chen and Narsavage (2006) examine the relationship among physiological, psychological and social factors and hospital readmission to develop a model predicting COPD hospital discharge.

3.1.2 Statistical Models for Binary Responses

In this research, we will consider binary readmission outcomes indicating whether the patient was readmitted within 30 days after discharge. Thus, statistical models for binary responses need to be found. The most popular model used in readmission studies is the logistic regression (LR) model, which is a special type of

Generalized Linear Model (GLM) commonly used for binary outcomes (Myers et al. 2002). Only a few studies consider special issues in model construction such as variable selection (Cao et al., 2006; Ferraris et al., 2001) and correlation in the outcomes (Hasan et al., 2007). Overall, the modeling of readmission is an underdeveloped field in healthcare studies. A comprehensive review of statistical models for binary responses is provided as follows.

Binary response is used when there are only two possible values a variable can take: 0 or 1, representing without or with readmission. The statistical models available for binary outcomes can be divided into two categories depending on whether the outcomes are independent or correlated. When the data come from different patients, that is, there is only one observation for each patient, the binary outcomes are assumed to be independent. In contrast, when there are more than one observation from the same patient on account of multiple readmissions, the outcomes are assumed to be correlated. It needs to be pointed out that when the correlation between outcomes from the same patients is believed to be moderate or the number of patients with multiple admissions is relatively small, the independence assumption will be applied to avoid the complexity in characterizing the correlation structure of the outcomes. After a complete search in the statistical literature, we find four major models for binary responses: the Logistic Regression (LR) model and the Logistic Regression Tree (LRT) model for independent outcomes, and the Generalized Estimating Equations (GEE) and the Generalized Linear Mixed Models (GLMM) for correlated outcomes. Basics of each model is given as follows.

3.1.2.1 Logistic Regression

In the Logistic Regression model (Myers et al., 2002), the binary response variable is assumed to follow a Binomial distribution

$$y_i \sim \mathrm{Binomial}(p_i)$$

Where y_i is the readmission outcome of patient i and p_i is the probability of readmission of patient i. This probability depends on the covariates through

$$\eta(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$$

Where x_1, x_2,…. are the risk factors, β_0, β_1,….are the parameters of this model which represent the effects of the risk factors on the probability of readmission. These parameters are usually estimated using maximum likelihood estimation methods. η is the link function which connects the mean of the response variable to the linear function of the risk factors in such a way that the range of the nonlinearly transformed mean ranges from -∞ to +∞. Some popular link functions used for GLMs are:

$$\text{cloglog link function} \quad \eta(p_i) = \ln(-\ln(1 - p_i))$$

$$\text{Logit link function} \quad \eta(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right)$$

$$\text{Probit link funtion} \qquad \eta(p_i) = \Phi^{-1}(p_i), \text{where } \Phi(\xi) = \frac{1}{\sqrt{2\prod}} \int_{-\infty}^{\xi} e^{\frac{-z^2}{2}} dz$$

The LR model is used extensively to model binary responses, especially in medical and social science studies. This model is simple and conceptually easy to understand. Moreover, due to the popularity of the LR model, software for model building and diagnostics, like R, and SPSS, has become available and widely used in practice.

3.1.2.2 Logistic Regression Tree

The Logistic Regression Tree model proposed by Chan and Loh (2004) is a tree structure extension of The LR model. Like other regression tree models, the LRT model divides the sample space into subspaces and then builds simple LR models within these subspaces

$$\log \frac{p_i}{1 - p_i} = \begin{cases} \boldsymbol{\beta}_{\mathrm{I}} \mathbf{X}_i & \text{if } \mathbf{X}_i \in \text{subspace I} \\ \boldsymbol{\beta}_{\mathrm{II}} \mathbf{X}_i & \text{if } \mathbf{X}_i \in \text{subspace II} \\ & \qquad \cdots \end{cases}$$

Usually categorical covariates are used for the partition and continuous covariates are used in fitting the LR models. For example, patients may be divided into groups by their gender and/or race, and one LR model is fitted for each group. The main advantage of LRT models is that it can overcome the interpretability issues of the LR model in the face of multi-collinearity, nonlinearity and interactions, without sacrificing estimation accuracy. Another advantage lies in the intuitive graphical representation of the model structure.

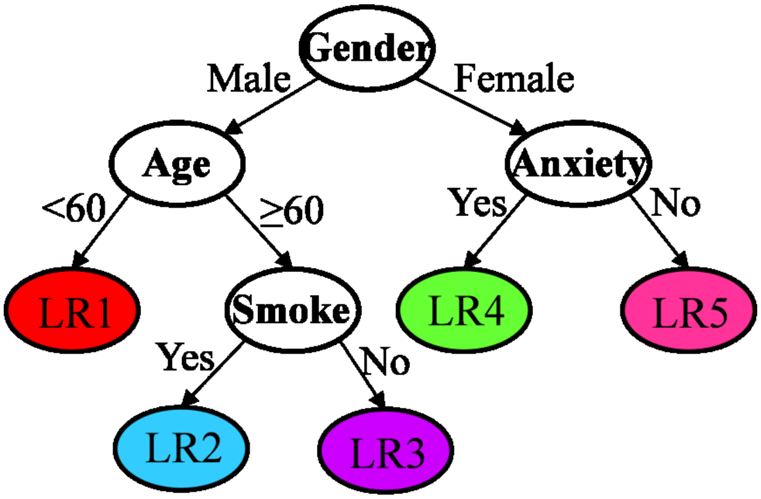


Figure 3-1 An Example of the Logistic Rregression Tree

Figure 3.1 shows an example of a Logistic Regression Tree model where the data space is divided into a number of spaces. For example, the first subspace contains patients who are male and younger than 60 years old, while the second subspace contains patients who are male, more than 60 years old and smoke regularly. An appropriate LR model is fitted for each of these subspaces. The LRT models fit the use in healthcare very well as it is a common practice to study the behaviors of subpopulations of patients in medical research.

3.1.2.3 Generalized Estimating Equations

The GEE model (Liang and Zeger, 1986) is an extension of generalized linear models which considers the correlation in the response data. It can estimate the effects of covariates more accurately in the presence of correlation. The GEE model takes the same model form as the LR model, except that the correlation among the data from same patients is taken into consideration. The GEE estimator of the model parameters can be obtained by solving

$$\mathbf{V} = A_i^{1/2}\mathbf{R}A_i^{1/2}$$
$$\sum_{i=1}^{n} \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}$$

where $\boldsymbol{\mu}_i$ is the mean vector which is a function of $\boldsymbol{\beta}$, **R** is a working correlation matrix of $\mathbf{y}_i$ which is common to all patients, **V** is the corresponding covariance matrix, and A_i is a $t \times t$ diagonal matrix with the variance of $\boldsymbol{\mu}_i$ as the diagonal elements. The working correlation matrix **R** needs to be specified for the estimation. Popular choices of **R** include:

Independence: the outcomes from same patients. This is based on the assumption that there is no correlation among the readmission outcomes of the same patients. In this case, the correlation takes the following form

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Exchangeable: This is also known as compound symmetry. In this model it assumes all the variances of readmission outcomes are equal and their pairwise covariances are also equal. This means that every observation is equally correlated with every other observation. Let *ρ* be the correlation coefficient between two outcomes, the corresponding correlation matrix is

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

Auto Regression 1 (AR1): In a time series analysis when the observations are correlated to their own past values through the number of lags between them then this phenomena is called auto-regressive. In an autoregressive correlation structure the two observations close to each other over time or space are more highly correlated than observations spreading further apart. The AR1 structure has homogenous variances and correlations that decline exponentially with distance. If ρ is the correlation coefficient then,

$$\mathbf{R} = \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

Unstructured: This is the most liberal structure, which means that there is no pattern at all. Each variance and each covariance is different and has no correlation to others. That is, the correlation matrix is of the following form

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

The GEE estimates can be obtained via the Newton-Raphson algorithm. One good property of this method is that it yields consistent estimates even when the correlation structure is mis specified. However, misspecification of the correlation structures will after the accuracy of the estimation.

3.1.2.4 Generalized Linear Mixed Models

The basic idea of the GLMM is to characterize the heterogeneity across patients by assuming the regression coefficients to be random and follow a certain probability distribution. Under this assumption, the outcomes of the same patients are correlated as they share the same unobserved coefficient. For this reason, the GLMM is often used to handle correlations in the outcomes. Specifically, the simplest form of the GLMM is

$$\log \frac{p_i}{1-p_i} = \boldsymbol{\beta}\mathbf{X}_i + u_i$$

where $\boldsymbol{\beta}$ is the fixed effect of covariates, and u_i is the random effect of patient i which are assumed to be independent and normally distributed as $N(0, \sigma^2)$. σ^2 is the variance of the random effect, which is a measure of the patient heterogeneity. In this model, the correlation of outcomes from the same patient is a constant depending on σ^2. The above GLMM is more precisely referred to as a random-intercept model as the random effect is on the intercept. The random effect can also occur on the coefficients of some predictors, e.g., β_k+δ, where β_k is the fixed effect of x_k, and δ is the associated random effect. The estimators of $\boldsymbol{\beta}$ and σ^2 can be obtained by the restricted maximum likelihood estimation method.

3.1.2.5 Selection of Models

For independent data, either the LR or the LRT can be used. Each has their pros and cons: The LR model is simple, but lacks of easy interpretation in complex cases where interactions between risk factors need to be considered; the LRT model bears better interpretability, but this good property will be affected when the model has many covariates and the tree structure is very complex. As is the case in any regression analysis, there is no “best” model for a given dataset, and it is useful to consider all possible ways of explaining the data and chose the best model through model comparison.

For correlated data, either the GEE model or the GLMM can be used. Generally speaking, the decision depends on the specific situation and the goal of the study. When the concern is the population averaged readmission, GEE should be used, while when the concern is heterogeneity among the patients in readmission, GLMM should be used (Hu et al., 1998). Another point is that in GLMM, the correlation of outcomes from the same patients is a constant for all patients, while in GEE there are other options. So GEE is able to characterize more complex correlation structures.

3.2 Problem Formulation

This study aims to develop a statistical model for patient readmission in COPD care based on a real data set provided by the University of Texas Medical Branch (UTMB) in Galveston, Texas.

Some important risk factors to COPD readmission are shown in Figure 3.2, including patient characteristics, e.g., age, gender, marital status and comorbidity, and process variables of the COPD care, e.g., the use of steroid and other treatments. The goal of this study is to build an appropriate statistical model to describe the dependency of readmission on the risk factors. The following issues need to be solved in this endeavor:

Determination of outcome correlation: We need to find whether the correlation among the readmission outcomes of the same patients is significant to determine which model to use. If the correlation is significant, then GEE or GLMM will be used; otherwise the LR or LRT model will be used.

Model selection: When the correlation of outcomes is determined, the best model for the given dataset needs to be found by comparing the candidate models. For example, if it is found that the outcomes can be viewed as independent, then the LR and LRT model will be compared to determine the one that fits the data better.

Variable selection: Significant risk factors to patient readmission need to be identified through variable selection techniques.

Link function selection: As given in section 3.1.2.1, there are three popular link functions for the GLMs. We need to determine which link function works the best for the data through comparing their predictive performance.

The proposed modeling approach which can solve the above issues is described in the following section.

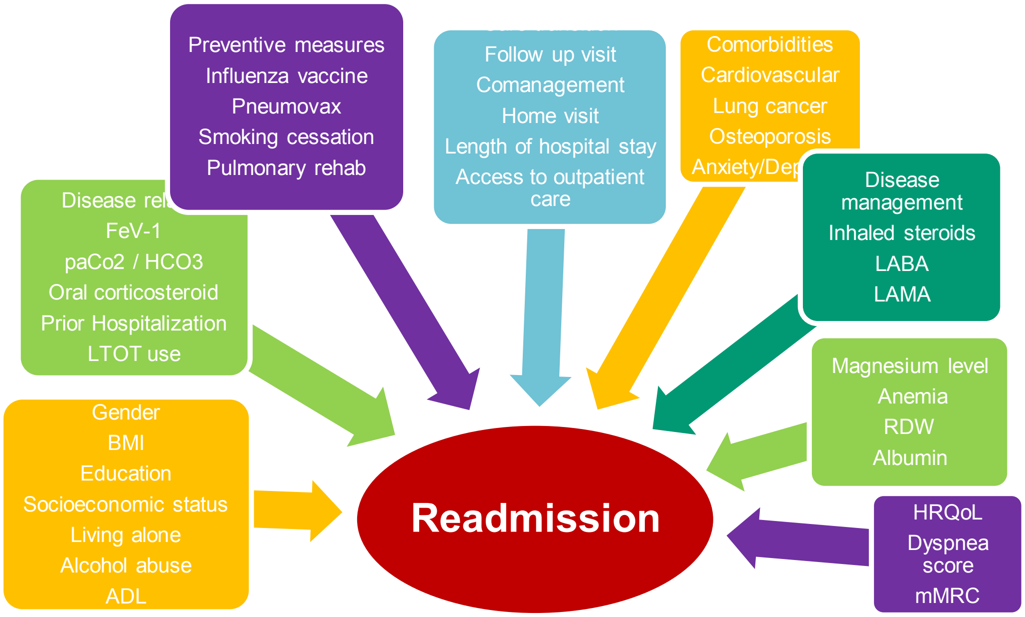


Figure 3-2 Risk factors to COPD readmission

3.3 The Proposed Approach

In this research we propose a systematic approach to build models for patient readmission. Figure 3.3 shows a schematic representation of our proposed approach. In the first step we determine whether the correlation among outcomes of the same patients is significant by fitting a Generalized Estimating Equations (GEE) model. The model estimates contain the estimate of the correlation. If the correlation estimate is not significant then we will treat the data as independent data; otherwise they will be treated as correlated data. For independent data, the LR and the LRT model will be built separately and compared through cross validation to determine the fitting model for data. Variable selection and link function selection will also be considered in building the two models. If the outcomes are correlated, the GEE model and the GLMM will be built and compared. The best model selected in this procedure will be used in the monitoring task

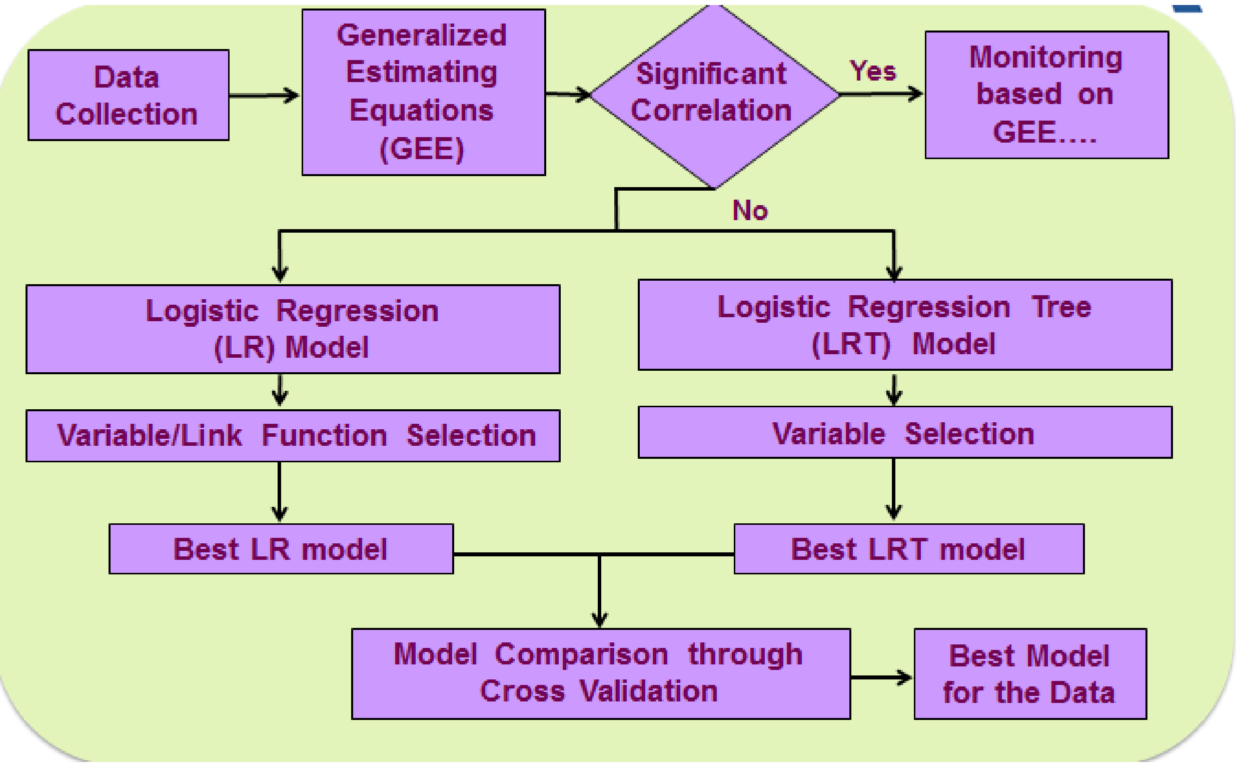

Data Collection
Generalized Estimating Equations (GEE)
Significant Correlation
Yes
Monitoring based on GEE….
No
Logistic Regression (LR) Model
Logistic Regression Tree (LRT) Model
Variable/Link Function Selection
Variable Selection
Best LR model
Best LRT model
Model Comparison through Cross Validation
Best Model for the Data

Figure 3-3 Overview of the Proposed Approach

The variable/link function selection will follow these steps:

(1) Fit a simple LR model for each individual covariate (main effect). The significant covariates are retained for the next step. The purpose of this screening is to identify all the potential significant covariates from the pool of available covariates.

(2) Fit a simple LR model for each two-way interaction of the selected covariates in last step. Significant interactions will be retained.

(3) Build a LR model using all the significant covariates and interactions through model selection techniques such as stepwise selection and likelihood ratio tests.

(4) Once the LR model is determined, comparing the three link functions in terms of their predictive performance and choose the best one.

(5) Build the LRT model using the selected covariates in the first step.

(6) Compare the LR and LRT model through cross validation.

The proposed approach will be applied to the dataset from COPD patients in a case study. The results of the analysis will be given in the following section.

3.4 Case Study

In this research the response variable is 30-day readmission of COPD patients from the University of Texas Medical Branch (UTMB), Galveston, TX. As shown in Figure 3.4, those patients were from all over the country but the majority of them were from Galveston, TX.

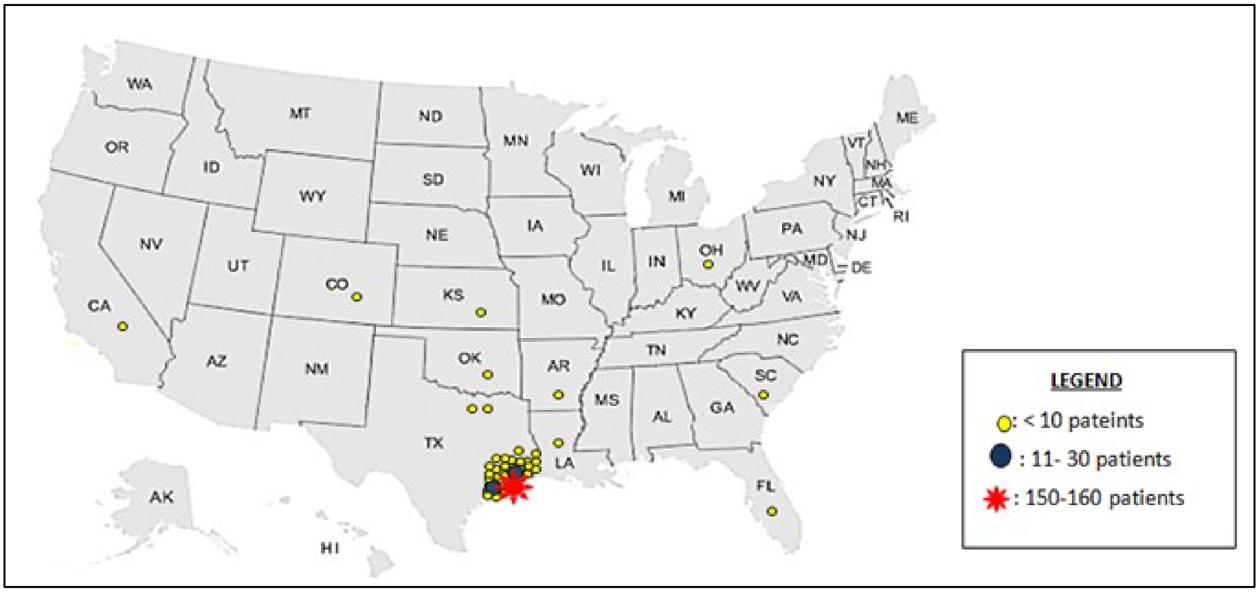


Figure 3-4 Spatial Distributions of the Patients

The data consist of the readmission data and 47 covariates as shown in Table 3.1. After data cleaning (removing “NA”), we have 282 observations.

Table 3-1 Variables in the Dataset

Variable number	Name
1	ALL
2	X48hrs
3	FOLLOWUP_15
4	FOLLOWUP_30
5	COPD_ORDERSET_USAGE
6	PFT_ORDERED_EVER
7	GENDER
8	MARITAL_STATUS
9	CS
10	CS_CURRENT_ORDER
11	LABA
12	LABA_CURRENT_ORDER
13	LAMA
14	LAMA_CURRENT_ORDER

Table 3-1 - *Continued*

15	ER_VISITS
16	OUPT_VISITS
17	HOSPITALIZATIONS
18	HOSPITALIZATIONS_COPD
19	OXYGEN
20	SMOKER
21	ALCOHOL_USE
22	DRUG_USE
23	DEPRESSION
24	ANXIETY
25	LUNG_CANCER
26	DIABETES
27	HYPERTENSION
28	CONGESTIVE_HEART_FAILURE
29	CORONARY_ARTERY_DISEASE
30	OSTEOPOROSIS
31	FLU
32	PNEUMOCOCCAL
33	PULMONARY_REHAB
34	PFT_ORDERED
35	ANTIBIOTICS_OVER1_DOSE
36	ANTIBIOTICS_ALL_DOSES
37	HEMOGLOBIN
38	RDW
39	EOS,
40	EOS_PERCENT
41	WBC
42	MAGNESIUM
43	LOS
44	Admission Source
45	RACE
46	AGE
47	FIN_CLASS

Among the patients, 78% of them have only one observation, 14.5% of them have two observations, and 7.5% of them have more than two observations. This means that most of the data are from different patients and thus can be assumed to be independent.

3.4.1 Univariate Analysis at alpha level 0.05

A univariate analysis is first done to select significant main effects, that is, a simple logistic regression model was fitted for each main effect of the risk factors separately. Significant main effects at α=0.05 are shown in Table 3.2.

Table 3-2 Significant Main Effects in the Univariate Analysis

Variable number	**Effects**	**p value**
1	LAMA	0.0142 *
2	ER_VISITS	3.14e-10 ***
3	HOSPITALIZATIONS	8.53e-07 ***
4	HOSPITALIZATIONS_COPD	0.00027 ***
5	OXYGEN	0.000284 ***
6	ALCOHOL_USE	0.045 *
7	DRUG_USE	0.0191 *
8	FLU	0.0231 *
9	RDW	0.00688 **

From the results in Table 3.2, we can see that the two most significant covariates are the number of previous emergency room visits (ER_VISITS) and the number of hospitalizations (HOSPITALIZATIONS). This finding is consistent with some existing studies on COPD readmission (e.g., Smith et al., 2000; Almagro et al., 2006; Bahadori and FitzGerald, 2007).

3.4.2 Fitting a Model for all the Selected Main Effects

The significant main effects shown in Table 3.2 were modeled using the LR and the GEE (link logit). The comparison of the p values in these two models is shown in Table 3.3.

Table 3-3 Comparison of Estimates of the GEE and LR Model

Model	Alpha level 0.001	Alpha Level 0.01	Alpha Level 0.05	Alpha Level 0.1
GEE model	ER_VISITS (1.4×10^{-05})	HOSPITALIZATIONS (0.00127) HOSPITALIZATIONS_COPD (0.003426)	LAMA (0.01697)	ALCOHOL_USE (0.054091)
Logistic Regression model	ER_VISITS (1.3×10^{-07})	HOSPITALIZATIONS (0.0064) HOSPITALIZATIONS_COPD (0.0071)	LAMA (0.0195) ALCOHOL_USE (0.0435)	

From the results in Table 3.3, we can see that the significance of the main effects is different in the two models. For example, the p value of ER-VISITS in the LR model is smaller than that in the GEE model, meaning that this factor is more significant in the LR model. Similarly, the factor ALCOHOL_USE has a smaller p value in the LR model. If α=0.05 is used, this factor will be significant in the LR model and not significant in the GEE model. This is actually consistent with the essential difference between the two models. In general, the GEE tends to degrade the significance of covariates because it takes the correlation in the outcomes into consideration which will increase the standard error in estimating the effect of each factor. As a result, some factors that are significant in the LR model will become insignificant when the GEE model is used.

3.4.3 Estimation of Correlation

The estimate of the correlation among readmission outcomes from same patients is as follows

```
Estimated Correlation Parameters using GEE:
Estimate Std.err
alpha   0.1602  0.1101
```

Number of clusters: 212 Maximum cluster size: 6

The Confidence Interval is given by equation

CI= estimate ± 1.96 (std. error) = [–0.0556, 0.3759]

Since the above confidence interval contains "0" in it, we determine that the correlation is not significant. Thus the data can be viewed as independent and the two models for independent outcomes, the LR and the LRT model, should be used.

3.4.4 Significant Two-Way Interaction at 0.05 Level

To select the significant two-way interactions, a univariate analysis is done to the interactions of the significant main effects. Again, a simple LR model is fitted to each interaction using a logit link. The following interactions are significant at α=0.05:

ER_VISITS:HOSPITALIZATIONS (p value = 0.0041 **)

ER_VISITS:DRUG_USE (p value = 0.028 *)

HOSPITALIZATIONS_COPD:OXYGEN (p value = 0.0437 *)

OXYGEN:RDW (p value = 0.0499 *)

3.4.5 Build the Complete LR Model

The complete Logistic Regression model (Link = logit) containing the significant main effects and two-way interactions is below. The estimates of the parameters and standard errors are displayed in Table 3.4.

X30DAY_READMISSION ~ LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + RDW + ER_VISITS:HOSPITALIZATIONS + ER_VISITS:DRUG_USE + HOSPITALIZATIONS_COPD:OXYGEN + OXYGEN:RDW

Table 3-4 Coefficients in the complete model

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-4.289	2.308	-1.859	0.0631 .
LAMA	1.220	0.502	2.428	0.0152 *
ER_VISITS	1.897	0.410	4.620	3.85×10^{-06}***
HOSPITALIZATIONS	0.176	0.072	2.444	0.0145 *
HOSPITALIZATIONS_COPD	0.377	0.279	1.352	0.1765
OXYGEN	-3.173	3.345	-0.949	0.3427
ALCOHOL_USE	0.782	0.429	1.821	0.0687 .
DRUG_USE	0.130	0.799	0.163	0.8702
FLU	0.285	0.455	0.627	0.5306
RDW	-0.026	0.152	-0.171	0.8640
ER_VISITS:HOSPITALIZATIONS	0.012	0.035	0.345	0.7303
ER_VISITS:DRUG_USE	-1.064	0.646	-1.647	0.0996.
HOSPITALIZATIONS_COPD:OXYGEN	-0.638	0.286	-2.227	0.0259*
OXYGEN:RDW	0.293	0.220	1.335	0.1820

3.4.6 Variable Selection

Two variable selection methods were used on the complete model: stepwise selection and likelihood ratio test. The likelihood ratio test starts from the full model and remove a covariate when the test is insignificant. Results in each step are shown in the following sections. Table 3.5 lists the results of the stepwise selection, where “mcomp” is the complete LR model found in the previous analysis.

Table 3-5 Results from Stepwise Selection

Model	Predictors	Deviance	df	AIC
mcomp	LAMA, ER_VISITS, HOSPITALIZATIONS, HOSPITALIZATIONS_COPD, OXYGEN, ALCOHOL_USE, DRUG_USE, FLU, RDW, ER_VISITS:HOSPITALIZATIONS, ER_VISITS:DRUG_USE, HOSPITALIZATIONS_COPD:OXYGEN ,OXYGEN:RDW	170.99	269	198.99
Xoxyrdw	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + RDW + ER_VISITS:HOSPITALIZATIONS + ER_VISITS:DRUG_USE + HOSPITALIZATIONS_COPD:OXYGEN	172.95	270	198.95
Xhoscopoxy	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + RDW + ER_VISITS:HOSPITALIZATIONS + ER_VISITS:DRUG_USE	177.12	271	201.12
Xervisianddrug	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + RDW + ER_VISITS:HOSPITALIZATIONS + HOSPITALIZATIONS_COPD:OXYGEN	175.33	271	199.33
Xervisinhosp	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + RDW + HOSPITALIZATIONS_COPD:OXYGEN	175.53	272	197.53
Xrdw	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + HOSPITALIZATIONS_COPD:OXYGEN	176.41	273	196.41
Xflu	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + HOSPITALIZATIONS_COPD:OXYGEN	177.1	274	195.1
Xdruguse	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + HOSPITALIZATIONS_COPD:OXYGEN	178.38	275	194.38
Xalcuse	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + HOSPITALIZATIONS_COPD:OXYGEN	182.01	276	196.01

Table-3.5 - *Continued*

Xoxy	LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + HOSPITALIZATIONS_COPD:OXYGEN	185.88	277	197.88
Xhospco	LAMA + ER_VISITS + HOSPITALIZATIONS + OXYGEN + HOSPITALIZATIONS_COPD:OXYGEN	183.66	277	195.66
Xhosp	LAMA + ER_VISITS + OXYGEN + HOSPITALIZATIONS_COPD:OXYGEN	196.71	278	206.71
Xerv	LAMA + HOSPITALIZATIONS + OXYGEN + HOSPITALIZATIONS_COPD:OXYGEN	220.99	278	230.99
Xlama	ER_VISITS + HOSPITALIZATIONS + OXYGEN + HOSPITALIZATIONS_COPD:OXYGEN	188.65	278	198.65

Table 3.6 lists the results from the likelihood ratio test. At each step one covariate was removed and the reduced model was compared to the previous model by the likelihood ratio test. If the p value of the test is less than 0.05, that means there is a significant difference between the previous (bigger) model and the reduced model. Thus the model selection will stop and the previous model was retained; otherwise the selection will proceed. Note that the likelihood ratio can only compare two nested models.

Table 3-6 Results from Likelihood Ratio Test

Models Compared	**Deviance difference**	**p value of LRT test**	**Decision**
mcomp vs Xoxyrdw	-1.961	0.1613	model without the interaction oxygen:RDW. Keep the model Xoxyrdw
Xoxyrdw vs Xhoscopoxy	-4.161	0.0413	p value less than 0.05 so we keep model with the hospitlaztion_COPD:oxygen. Reject the model Xhoscopoxy and keep the model Xoxyrdw

Table 3.6 - *Continued*

Xoxyrdw vs Xervisianddrug	-2.374	0.1233	model without the interaction ER_VISITS:DRUG_USE. Keep the model Xervisianddrug
Xervisianddrug vs Xervisinhosp	-0.201	0.6531	model without the interaction ER_VISITS:HOSPITALIZATIONS. Keep the model Xervisinhos
Xervisinhosp vs Xrdw	-0.875	0.3494	model without RDW. Keep the model Xrdw
Xrdw vs Xflu	-0.697	0.4037	model without FLU. Keep the model Xflu
Xflu vs Xdryguse	-1.280	0.2578	model without DRUG_USE. Keep the model "Xdruguse"
Xdruguse vs Xalcuse	-3.631	0.0566	model without ALCOHOL_USE. Keep the model "Xalcuse"
Xalcuse vs Xoxy	-3.866	0.0492	p value less than 0.05 so we keep modle with the oxygen. Reject the model Xoxy and keep the model Xalcuse.
Xalcuse vs Xhospco	-1.649	0.199	model without HOSPITALIZATIONS_COPD. Keep the model Xhospco
Xhospco vs Xhos	-13.041	0.0003	p value less than o.o5 so we keep the model with HOSPITALIZATIONS. Reject the model Xhos and keep the model Xhospco
Xhospco vs Xerv	-37.321	1×10^{-09}	p value less than 0.05 so we keep the model with ER_VISITS. Reject the model Xerv and keep the model Xhospco
Xhospco vs Xlama	-4.987	0.0255	p value less than 0.05 so we keep the model with LAMA. Reject the model Xlama and keep the model Xhospco

3.4.7 Comparing Different Link Functions

The final logistic regression model is then compared by changing the link functions. The three link functions are logit, probit and cloglog. The receiver operating characteristic (ROC)curves were made for link function as shown in Figures 3.5, 3.6 and 3.7.

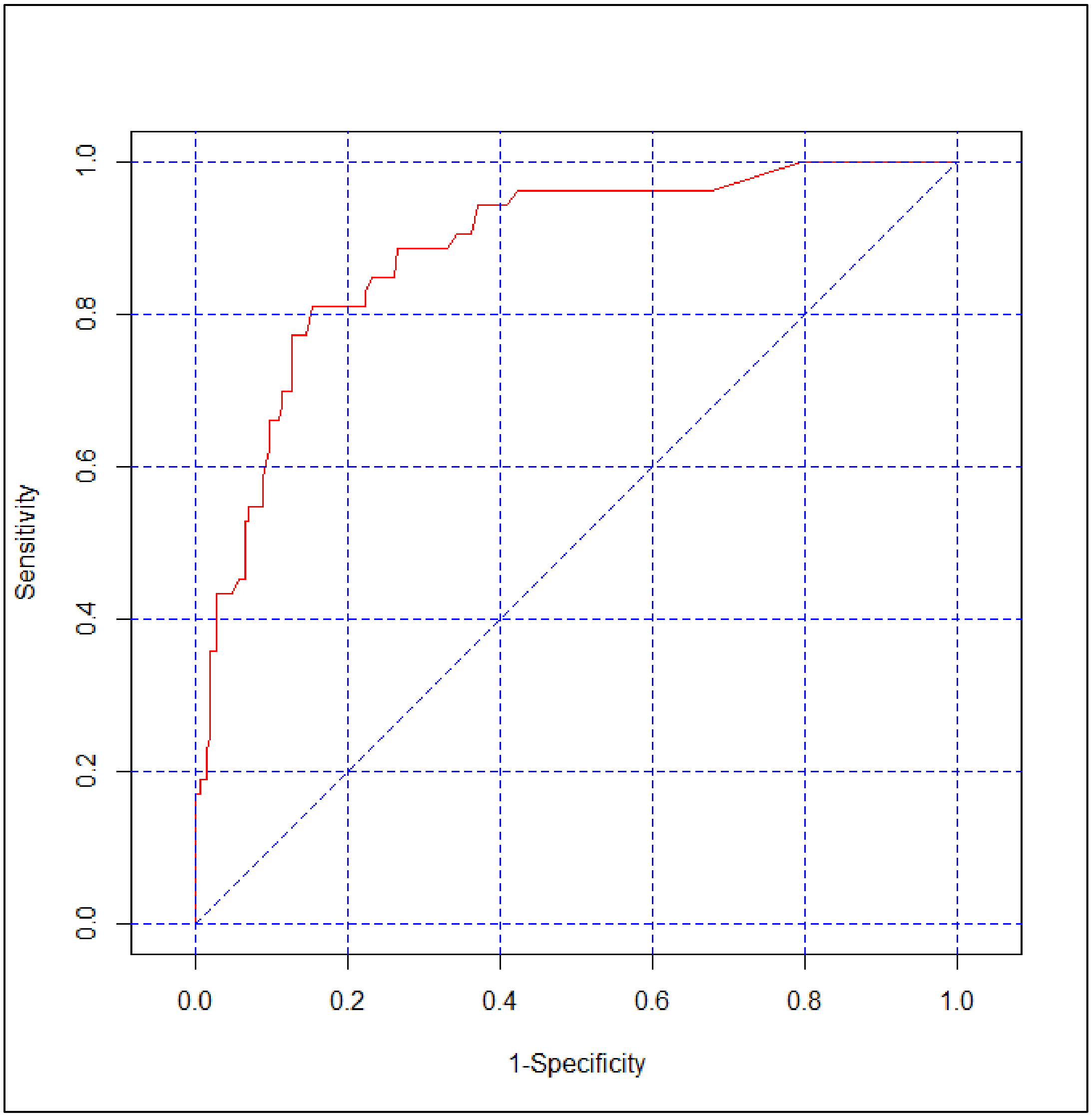


Figure 3-5 ROC Curve for the Logit Link Function

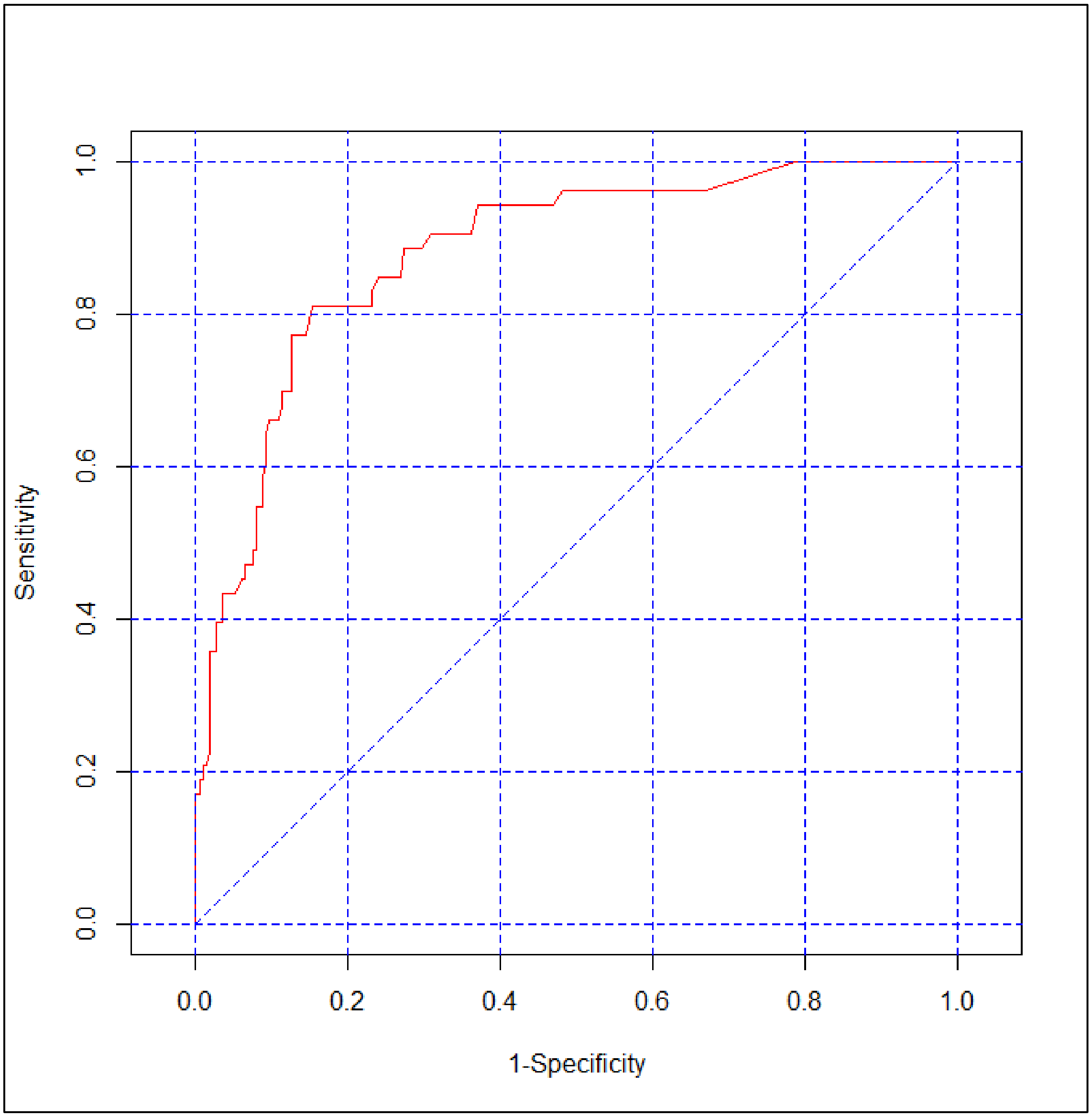


Figure 3-6 ROC Curve for the Probit Link Function

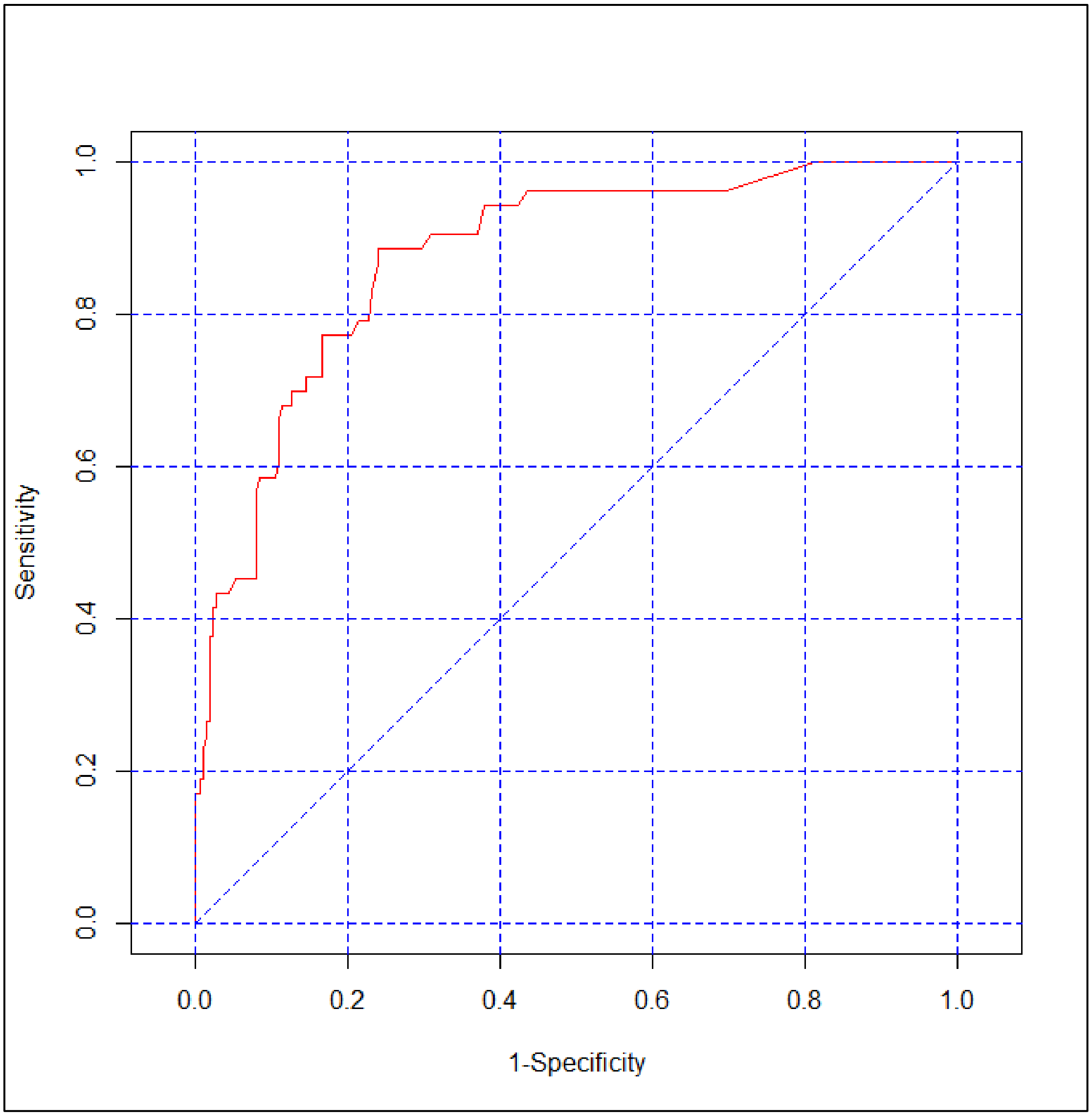


Figure 3-7 ROC Curve for the Cloglog Link Function

3.4.8 Performance Criteria for Selecting the Best LR Model

Different performance criteria can be applied in selecting the best LR model which represent different perspectives to evaluate the models. For example if the prediction performance is of interest, then criteria on how well the model can predict future values should be used. If the fitting performance is concerned, then criteria on how well the model fits the data need to be used. Some of the commonly used model selection criteria are:

AIC (Akaike Information Criterion)

BIC (Bayesian Information Criterion)

Prediction- Cross Validation or Area under ROC curve (AUC)

Simplicity

The final model from the stepwise selection and backward elimination based on the AIC criteria is:

MODEL-backstep: X30DAY_READMISSION ~ LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + ER_VISITS:DRUG_USE + HOSPITALIZATIONS_COPD:OXYGEN)

The final model from the forward model selection is:

MODEL-fwdstep**:** X30DAY_READMISSION ~ LAMA + ER_VISITS + HOSPITALIZATIONS + HOSPITALIZATIONS_COPD + OXYGEN + ALCOHOL_USE + DRUG_USE + FLU + RDW + ER_VISITS:DRUG_USE + HOSPITALIZATIONS_COPD:OXYGEN + OXYGEN:RDW)

The results from the stepwise selection and backward selection are the same. However, the results from the forward selection are different. Since a lower AIC value indicates a better fitting of the model, the model selected by the backward selection is better than that from the forward selection.

The final model from the backward elimination based on the p value criteria is:

MODEL-Backlrt: X30DAY_READMISSION ~ LAMA + ER_VISITS + HOSPITALIZATIONS + OXYGEN + HOSPITALIZATIONS_COPD:OXYGEN)

Another performance criterion is the Area under the ROC curve (AUC). In the following section the final models will be compared according to their AUCs. Since link functions can also play a significant role, we have considered all the three link function for each model.

Table 3-7 Performance Evaluation for Various LR Models

MODEL	Link Function	AIC	Residual Deviance	AUC
Backstep_logit	Logit	194.72	174.72	88.73%
Backstep_probit	Probit	193.67	173.67	88.79%
backstep_cloglog	cloglog	198.01	178.01	88.13%
fwd_logit	Logit	197.12	171.12	88.43%
fwd_probit	Probit	197.01	171.01	88.55%
fwd_cloglog	cloglog	199.41	173.41	88.02%
Backlrt_logit	Logit	195.66	183.66	88.42%
Back lrt_probit	Probit	193.92	181.92	88.23%
Back lrt_cloglog	cloglog	201.37	189.37	87.76%

From the results in Table 3.7, it can be seen that all the link functions deliver similar results. Hence we can use the most popular link function, the logit link function. Since the backstep model that was found using the stepwise model selection has the best AUC, the best LR model is:

BEST MODEL: X30DAY_READMISSION ~ -4.49 + 1.15 ×x LAMA + 1.96 x ER_VISITS + 0.19xHOSPITALIZATIONS + 0.33xHOSPITALIZATIONS_COPD + 1.05xOXYGEN + 0.76 xALCOHOL_USE + 0.05 x DRUG_USE -0.91 x ER_VISITS:DRUG_USE - 0.57 xHOSPITALIZATIONS_COPD:OXYGEN)

3.4.9 Analytical Techniques for Predictive Analysis and Logistic Regression Tree (LRT)

Predictive analysis uses machine learning, modeling, statistics and data mining to analyze the current scenario using historical facts to make predictions about future. The approaches and techniques to carry predictive analysis are:

- Regression models
- Linear Regression models

- Discrete choice models
- Logistic regression
- Time series models
- Survival or duration analysis
- Classification and Regression Trees (CART)
- Multivariate Adaptive Regression Splines (MARS)
- Logistic Regression Tree (LRT)
- Machine learning techniques
- Neural networks.

Classification and Regression Trees (CART), Multivariate Adaptive Regression Splines (MARS) and Logistic Regression Tree (LRT) are non-parametric decision tree learning technique. A decision tree predictive model plots observations about an item to draw conclusions about the item's target value. In the decision tree structure the leaves represent class labels and branches represent aggregations of features that lead to those class labels. A decision tree makes various rules based on the variables in the dataset. These rules are based on variable's values selected to get the split to differentiate observations based on dependent variable. Once a rule is selected and splits a node into two then there is a recursive procedure that is applied to each child node. Figure 3.8 shows a schematic comparison between a regular model with a tree model. Tree models have some advantages over the regular models such as

Tree models can be used to model more complex models.:Tree models are visually intuitive and convenient to use. Tree model does not have interactions which makes it easy to interpret.

They can provide important insights based on experts describing a situation.

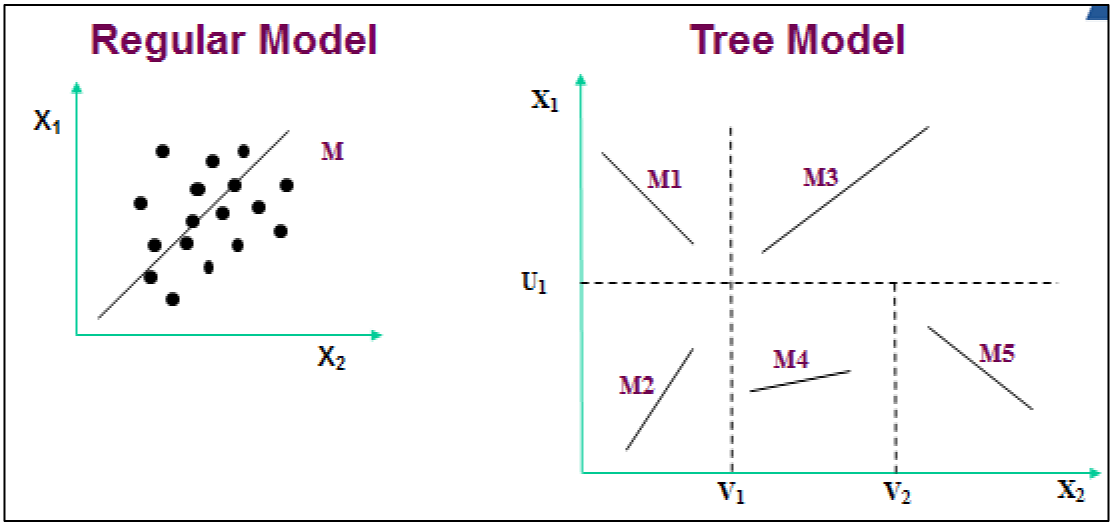


Figure 3-8 Comparison Between Regular Model and Tree Model

The Logistic Tree with Unbiased Selection (LOTUS) algorithm developed by Chan and Loh (2004) can be used to build the LRT model. The selected main effects of risk factors in section 3.4.1 will be used in fitting the LRT model. Those main effects are:

1. LAMA (Binary)
2. ER Visit (Numeric)
3. Hospitalizations (Numeric)
4. Hospitalization_COPD (Numeric)
5. Oxygen (Binary)
6. Alcohol use (Binary)
7. Drug Use (Binary)
8. RDW (Numeric)
9. Flu (Binary)

In this study three ways to fit the LRT model are considered which assign different variables for splitting the tree and fitting the LR model at each end of the branch. The designations of LOTUS are shown in Figure 3.9.

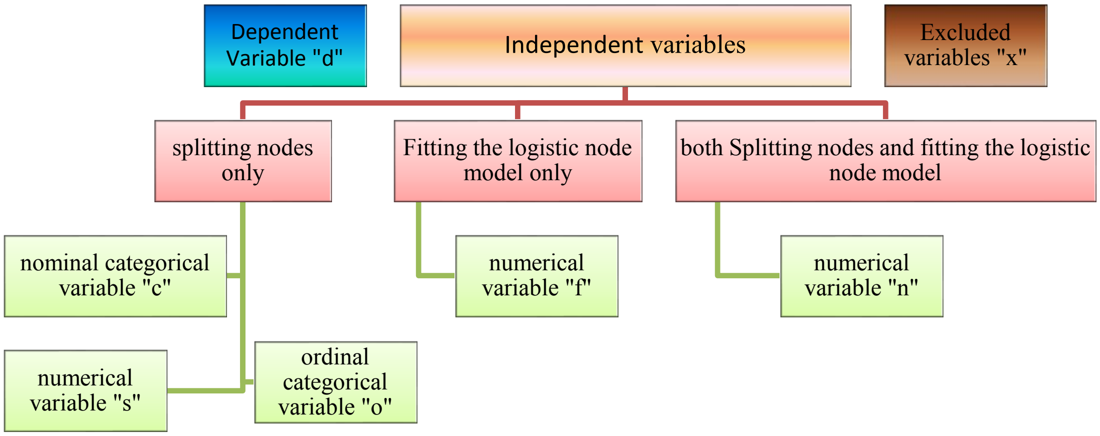

Dependent Variable "d"
Independent variables
Excluded variables "x"
splitting nodes only
Fitting the logistic node model only
both Splitting nodes and fitting the logistic node model
nominal categorical variable "c"
numerical variable "f"
numerical variable "n"
numerical variable "s"
ordinal categorical variable "o"

Figure 3-9 9 Designations of Variables in LOTUS

The explanation of each type of variable is as described below:

- Nominal categorical variable: A variable that has categories but the order is not important. For example Gender.
- Ordinal categorical variable: A variable that has categories and the order has some significance. For example, financial status Low, medium and high.

LRT Model 1: The designation of variables in this model is shown in Table 3.8. All the categorical variables are designated by "c", which means that they are used for splitting nodes only. All the numeric variables are designated by "f", which means that they are used for fitting the logistic model only.

Table 3-8 Designation of variables in LRT Model 1

S/NO	**Variable**	**Variable type**	**LOTUS denotation**
Response varibale	30 day readmission	categorical (binary)	d
1	LAMA	categorical (binary)	c

Table 3.8- *Continued*

2	ALCOHOL_USE	categorical (binary)	c
3	ER_VISITS	numeric	f
4	HOSPITALIZATIONS	numeric	f
5	HOSPITALIZATIONS_COPD	numeric	f
6	OXYGEN	categorical (binary)	c
7	DRUG_USE	categorical (binary)	c
8	FLU	categorical (binary)	c
9	RDW	numeric	f

The Logistic Regression Tree model diagram for the designation given in Table 3.8 is shown in Figure 3.10.

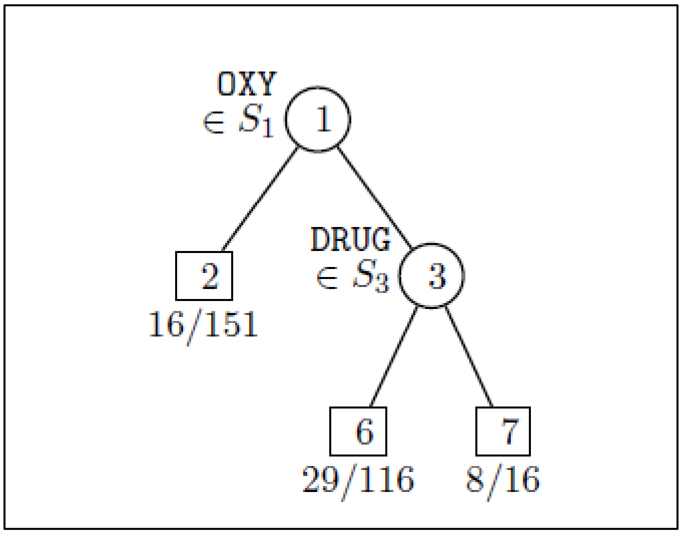


Figure 3-10 Logistic Regression Tree Model 1

Logistic regression tree output

Regression tree output:

Node 1: OXY = 0

Node 2: Probability = 0.1060E+00

Node 1: OXY = 1

Node 3: DRUG = 0

Node 6: Probability = 0.2500E+00

Node 3: DRUG = 1

Node 7: Probability = 0.5000E+00

Terminal node models of logistic regression tree

Node 2: Deviance = 7.9762E+01

Total Cases = 151, Cases Fit = 151

Total Cases with Y=1 = 16

Variable	Coefficient	Std Error	T-Value
Intercept	-3.815	6.259×10^{-01}	-6.094
ER	1.931	5.037×10^{-01}	3.833
HOSC	6.313×10^{-01}	2.569×10^{-01}	2.457

Model at terminal node 2 is:

$(OXY=0) \sim -3.815 + 1.931 \times ER + 6.3134 \times HOSC + (6.259 \times 10^{-01} + 5.037 \times 10^{-01} + 2.569 \times 10^{-01})$

Node 6: Deviance = $8.715 \times 10^{+01}$

Total Cases = 116, Cases Fit = 116

Total Cases with Y=1 = 29

Variable	Coefficient	Std Error	T-Value
Intercept	-3.168	6.071×10^{-01}	-5.219
ER	1.809	4.704×10^{-01}	3.846
HOS	3.261×10^{-01}	1.118×10^{-01}	2.916
HOSC	-3.286×10^{-01}	1.501×10^{-01}	-2.189

Model at terminal node 6 is :

(OXY=1, DRUG = 0) ~ -3.168 + 1.809×ER +3.261×HOS – 3.286×HOSC + (6.071×10^{-01} + 4.704×10^{-01} + 1.118×10^{-01} + 1.501×10^{-01})

Node 7: Deviance = 1.3955E+01

Total Cases = 16, Cases Fit = 16

Total Cases with Y=1 = 8

Variable	Coefficient	Std Error	T-Value
Intercept	-1.466	6.597	-2.223
RDW	9.365×10^{-01}	4.223×10^{-01}	2.217

Model at terminal node 7 is :

(OXY=1, DRUG = 1) ~ -1.4668 + 9.365×10^{-01}×RDW + (6.5976 + 4.223×10^{-01})

LRT Model 2: The designation of variables in this model is given in Table 3.9. In this model the categorical variables are nominal and they are used for splitting nodes only. All the numeric variables are designated by “n” which means that they are used both for splitting the nodes and fitting the logistic node model.

Table 3-9 Designation of Variable in LRT Model 2

S/NO	**Variable**	**Variable type**	**LOTUS denotation**
Response varible	30 day readmission	categorical (binary)	d
1	LAMA	categorical (binary)	c
2	ALCOHOL_USE	categorical (binary)	c
3	ER_VISITS	numeric	n
4	HOSPITALIZATIONS	numeric	n

Table 3.9 – *Continued*

5	HOSPITALIZATIONS_COPD	numeric	n
6	OXYGEN	categorical (binary)	c
7	DRUG_USE	categorical (binary)	c
8	FLU	categorical (binary)	c
9	RDW	numeric	n

The logistic regression tree model diagram for the designation given in Table 3.9 is shown in Figure 3.11.

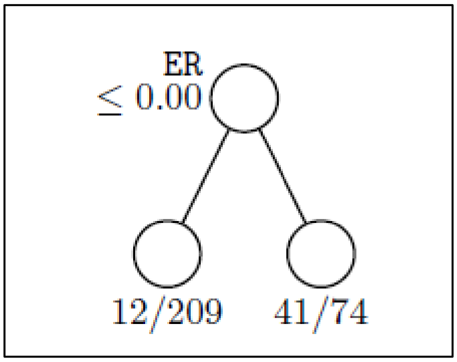


Figure 3-11 Logistic Regression Tree Model 2

Regression tree:

Node 1: ER <= 0.0000E+00

Node 2: Probability = 0.5742E-01

Node 1: ER > 0.0000E+00

Node 3: Probability = 0.5541E+00

Terminal Node Models of Logistic Regression Tree:

Node 2: Deviance = 8.6085E+01

Total Cases = 209, Cases Fit = 209

Total Cases with Y=1 = 12

Variable	Coefficient	Std Error	T-Value
Intercept	-3.631	5.070×10^{-01}	-7.162
HOS	1.670×10^{-01}	6.596×10^{-01}	2.531

Model at terminal node 2 is :

$(ER<=0) \sim -3.6319 + 1.67\times10^{-01}\times HOS+ (5.070\times10^{-01} + 6.596\times10^{-02})$

Node 3: Deviance = 9.4014E+01

Total Cases = 74, Cases Fit = 74

Total Cases with Y=1 = 41

Variable	Coefficient	Std Error	T-Value
Intercept	-5.833	2.409	-2.421
RDW	4.056×10^{-01}	1.623×10^{-01}	2.498

Model at terminal node 3 is :

$(ER >0) \sim -5.8338 + 4.056\times10^{-01}\times RDW + (2.409 + 1.623\times10^{-01})$

LRT Model 3: The designation of variables in this model is given in Table 3.10. In this model the categorical variables are nominal and they are used for splitting nodes only. All the numeric variables are designated by “f” which means that they are used for fitting the logistic node model only. The numerical variable ER visit is designated “n” which means that it is used both for splitting the nodes and fitting the logistic node model.

Table 3-10 Designation of Variables in LRT Model 3

S/NO	**Variable**	**Variable type**	**LOTUS denotation**
Response variable	30 day readmission	categorical (binary)	d
1	LAMA	categorical (binary)	c
2	ALCOHOL_USE	categorical (binary)	c
3	ER_VISITS	numeric	n
4	HOSPITALIZATIONS	numeric	f
5	HOSPITALIZATIONS_COPD	numeric	f
6	OXYGEN	categorical (binary)	c
7	DRUG_USE	categorical (binary)	c
8	FLU	categorical (binary)	c
9	RDW	numeric	f

The logistic regression tree model diagram for the designation given in Table 3.10 is shown in Figure 3.12.

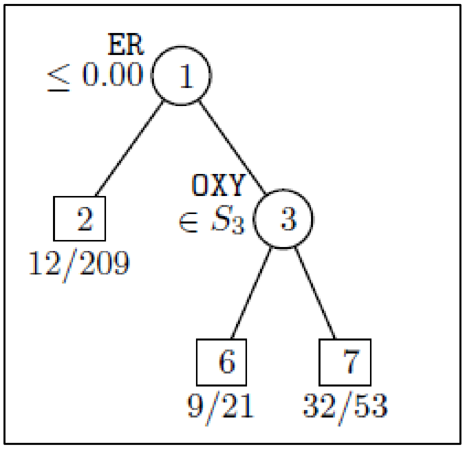


Figure 3-12 Logistic Regression Tree Model 3

Regression tree:

Node 1: ER <= 0.0000

Node 2: Probability = 0.574×10^{-01}

Node 1: ER > 0.0000

Node 3: OXY = 0

Node 6: Probability = 0.428

Node 3: OXY = 1

Node 7: Probability = 0.603

Terminal Node Models of Logistic Regression Tree:

Node 2: Deviance = 8.6085E+01

Total Cases = 209, Cases Fit = 209

Total Cases with Y=1 = 12

Variable	Coefficient	Std Error	T-Value
Intercept	-3.631	5.070×10^{-01}	-7.162
HOS	1.670×10^{-01}	6.596×10^{-02}	2.531

Model at terminal node 2 is :

$(ER <=0) \sim -3.6319 + 1.67\times10^{-01}\times HOS + (5.070\times10^{-01} + 6.596\times10^{-02})$

Node 6: Deviance = 2.0594E+01

Total Cases = 21, Cases Fit = 21

Total Cases with Y=1 = 9

Variable	Coefficient	Std Error	T-Value
Intercept	-3.273	1.348	-2.427
HOSC	1.937	8.570×10^{-01}	2.260

Model at terminal node 6 is :

$(ER >0, OXY=0) \sim -3.273 + 1.937\times HOS + (1.3484 + 8.570\times10^{-01})$

Node 7: Deviance = 6.3142E+01

Total Cases = 53, Cases Fit = 53

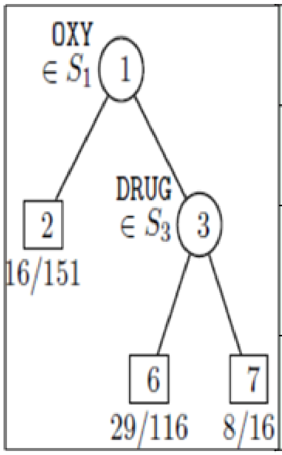
Total Cases with Y=1 = 32

Variable	Coefficient	Std Error	T-Value
Intercept	-7.635	3.287	-2.322
RDW	5.319×10^{-01}	2.196×10^{-01}	2.421

Model at terminal node 7 is :

(ER >0, OXY=1) ~ $-7.635 + 5.319\times10^{-01}\times\text{RDW} + (3.287 + 2.196\times10^{-01})$

The LRT Model 1 is chosen here for its intuitive interpretation. The model is shown in Figure 3.13.



Terminal Node	Model
2 (OXY=0)	$\log p/(1-p) =$ $-3.815+1.93\times$**ER** $+6.31\times$**HOSC**
6 (OXY =1, DRUG =0)	$\log p/(1-p) =$ $-3.16+1.8\times$**ER**$+3.2\times$**HOS**$-3.28\times$ **HOSC**
7 (OXY =1, DRUG=1)	$\log p/(1-p) =$ $-1.46+9.36\times10^{-1}\times$**RDW**

Figure 3-13 Fitted LRT model (left) and LR models at End Nodes (right)

3.4.10 Model Comparison

Cross validation technique is a model validation technique for evaluating how the results of a statistical analysis will generalize to an independent data set. Cross validation involves dividing a data set into complementary sub datasets. The analysis is performed on data set called the training data set and validated on another data set called testing data set. Cross validation is used when the goal is prediction and it is required to estimate how well the data will predict in a real situation. Common types of Cross validation are:

- K-fold cross validation
- 2-fold cross validation
- Repeated sub sampling validation
- Leave-one-out Cross Validation

In our study a simple 2-fold cross validation is applied. For each fold there is a random assignment of the data points so that each data set is of equal size. The advantage of 2-fold cross validation is that the training and testing data set are both large. In our case we randomly chose 80% data for training and 20 % data for testing.

Performance could be measured on the training data set using a threshold. A threshold is used to determine the predicted readmission. The rule is: if the predicted probability of readmission based on the considered model is larger than the threshold, then the prediction of the readmission is 1; otherwise, the prediction is 0. The threshold can be selected between 0 and 1. The success rate of a model is defined to be the percentage of simulations in which the prediction equals to the observation. In this study the cross validation is carried out for LR as well as LRT for each combination of threshold and number of simulations. Figure 3.14 shows the success rates of the two models for 100 simulations at a threshold of 0.5. Figure 3.15 shows the results for 10,000 simulations at a threshold of 0.7.

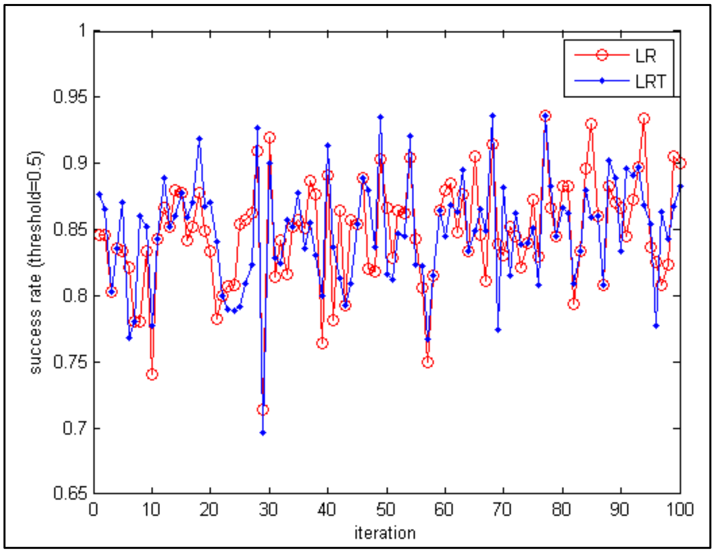


Figure 3-14 Success Rates Based on 100 Simulations at a Threshold of 0.5

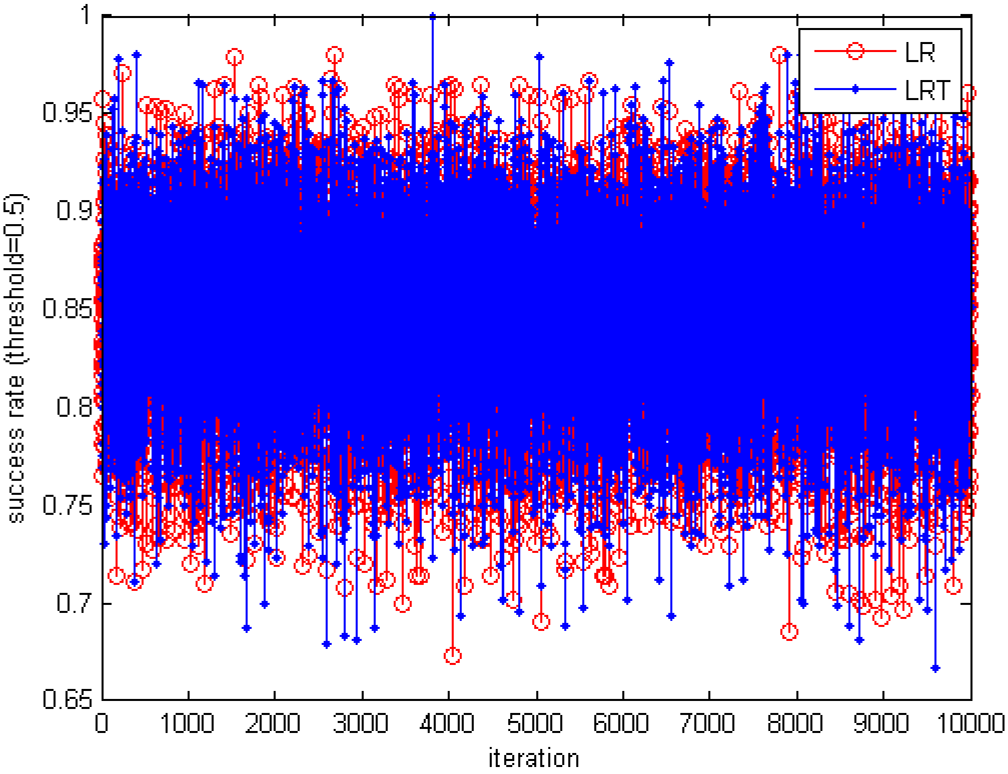


Figure 3-15 Success Rates Based on 10,000 Simulations at a Threshold of 0.7

From the figures above, it can be seen that both LR and LRT have similar results for performance. Table 3.11 shows the results of all the simulations carried at different thresholds.

Table 3-11 Simulation Results at Different Thresholds

Simulation number	Number of Simulations	Threshold	% simulations where LR performed better than LRT
1	10000	0.5	0.3623
2	10000	0.55	0.465
3	10000	0.6	0.4992
4	10000	0.65	0.4798
5	10000	0.7	0.4435
6	10000	0.75	0.3796
7	10000	0.8	0.3132
8	10000	0.85	0.2674
9	10000	0.9	0.2571
10	10000	0.95	0.2871

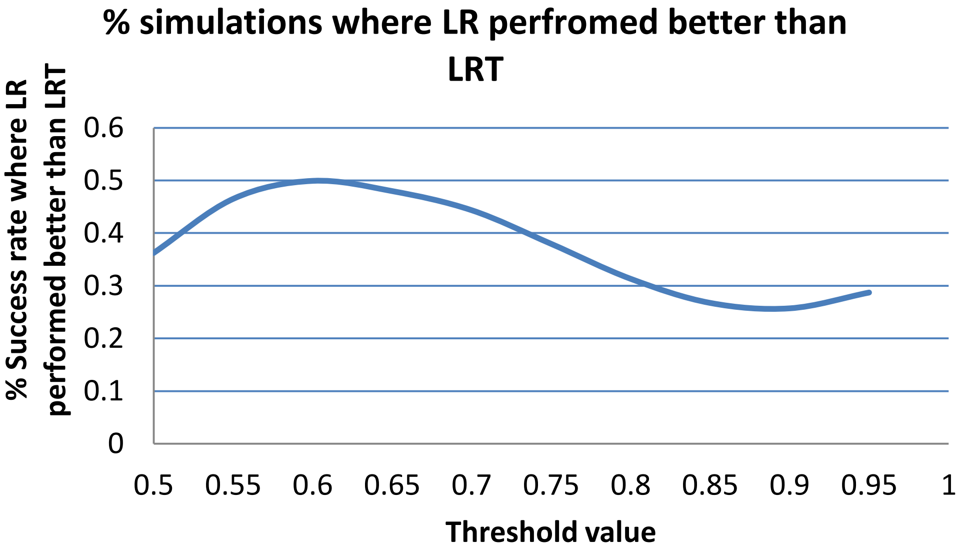


Figure 3-16 Comparison of Performance of LR vs LRT

Figure 3.16 shows the graphical summary of the percent simulations where LR performed better than LRT. From the figure we can see that LRT performs better than LR over 50% of the time under different threshold values. Hence we conclude that the LRT is the best model for the data. Risk-adjusted monitoring will be conducted based on this model in our future study.

Chapter 4

Risk-Adjusted Phase I Monitoring of Patient Readmission in COPD Care

4.1 Literature Review

Risk-adjusted Monitoring is very important in healthcare industry to ensure homogeneity among all the cases considered by the healthcare provider. Figure 4.1 shows a schematic diagram of the risk-adjusted monitoring in healthcare. In a non-risk-adjusted monitoring the patient outcome such as survival rate, readmission, adverse events etc. is a function of Quality of care only. However, unlike manufacturing processes where products are homogeneous, in health care scenario patients come from different backgrounds with various risk factors such as severity, comorbidity, age, etc., associated with them. Hence in a risk-adjusted monitoring patient outcome is a function of healthcare quality as well as the risk factors associated with the patient.

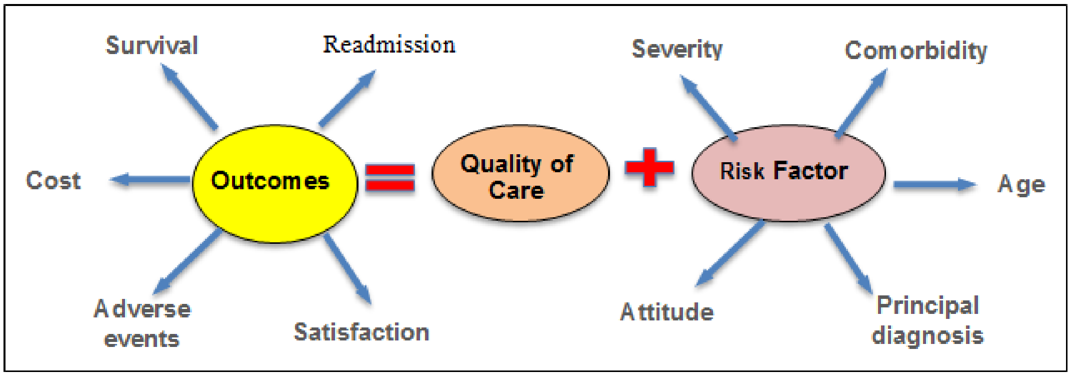


Figure 4-1 Risk-adjusted Monitoring in Healthcare

Many studies have been done on risk-adjusted monitoring of patient outcomes and Phase I monitoring. A brief review of literature on these two topics is given as follows.

4.1.1 Risk-adjusted Monitoring

Grigg and Farewell (2004), Woodall (2006), and Cook et. al (2008) give excellent reviews on methods and techniques for risk-adjusted monitoring of healthcare provider's performance. These methods can be divided into three categories: simple risk-adjusted plots, extension of non risk-adjusted SPC control charts, and Bayesian approaches. The proposed approaches in the first two categories focus on Phase II monitoring, while Bayesian approaches can be used for both Phase I and Phase II monitoring.

Simple risk-adjusted plots: Simple plots of the cumulative difference between observed and expected outcomes have been used to detect changes in surgical performance

$$C_t = C_{t-1} + (y_t - \mathrm{p}_t)$$

where C_t is the statistic at time t, y_t is the observed outcome (death/survival), and p_t is the baseline mortality probability of patient t. Obviously, if there is a sustained change in the performance of care providers, an increasing/decreasing trend can be seen in the plot. Such methods include the observed-expected (O-E) plot (Polonieki, 1998) and the variable life- adjusted (VLAD) plot (Lovegrove, et al., 1997). While these plots are very easy to implement and understand by healthcare practitioners, the statistical properties of the statistic monitored are not clear and thus it is very difficult to set up control limits.

Extension of non-risk-adjusted SPC control charts: As SPC is a well-studied area in other contexts especially industrial applications, various non-risk-adjusted control charts are available in the literature. These techniques have been extended to medical applications by incorporating risk adjustment. Popular extensions of these techniques include

(1) *Risk-adjusted p chart*: *p*-chart is a basic SPC technique to monitor binary data such as defectiveness or nondefectiveness of products in industrial process control where defective rate is an important concern. To apply this technique in Phase II monitoring, subgroups of data need to be collected, and a normal distribution is assumed when the sample size of subgroups is adequately large

$$\hat{p}_i = \frac{\sum_{t=1}^{n_i} y_{it}}{\mathrm{n_i}} \sim \mathrm{N}\left(\mathrm{p_0}, \frac{\mathrm{p_0}(1-\mathrm{p_0})}{\mathrm{n_i}}\right)$$

where $\mathrm{n_i}$ is the sample size of subgroup *i*, $\hat{p}_i$ is the corresponding average defective rate, $\mathrm{y_{it}}$ is the measurement of product in subgroup *i*, and p_0 is the base-line defective rate estimated from historical data. 3-sigma control limits can be obtained based on this distribution. To extend this method to medical contexts, patients in consecutive time periods of same length, e.g., 6 months, are grouped, and the distribution used in the monitoring becomes

$$\hat{p}_i = \frac{\sum_{t=1}^{n_i} \mathrm{y_{it}}}{\mathrm{n_i}} \sim \mathrm{N}\left(\frac{1}{\mathrm{n_i}}\sum_{\mathrm{t=1}}^{\mathrm{ni}} \mathrm{p_{0ti}}, \frac{1}{\mathrm{n_i}}\sum_{\mathrm{t=1}}^{\mathrm{ni}} \mathrm{p_{0ti}}(1-\mathrm{p_{0ti}}\right)$$

where $\mathrm{p_{0ti}}$ is the base-line mortality probability of patient *t* in the group *i*. Cockings, Cook and Iqbal (2006) and Cook, et al. (2003) use such charts to monitor mortality in intensive care. The risk-adjusted p-chart is easy in implementation and interpretation, and also provides a convenient way to set up control limits. However, the need of grouping patients in considerably long periods may lead to delay in capturing changes in performance.

(2) *Risk-adjusted set method*: The set method monitors the time between adverse events (e.g., death) by counting the number of events (e.g., survival) between

any two consecutive occurrences of such events. Specifically, letting C_t be the current set number, i.e., count of events following the occurrence of an interested event, $C_t = C_{t-1} + 1$, that is, this statistic will increase by 1 if the t^{th} observation is not the interested event. This continues until an interested event occurs, and then the set number will be reset to 0. An alarm is signaled when $C_t \leq T$ happens n times, where (T,n) is a pair of thresholds determined through simulation.

An extension of this method to incorporate risk-adjustment has been proposed by Grigg and Farewell (2004). The basic idea is to weigh each event by the base-line mortality probability of the patient. Specifically, the set number will be calculated by

$$C_t = C_{t-1} + \frac{p_{ot}}{\overline{p_0}}$$

where, p_{ot} is the base-line mortality probability of patient *t*, and $\overline{p_0}$ is the average base-line mortality probability of all patients, which can also be termed as the base-line mortality probability of an “average” patient. Here the average patient is used as a benchmark to assess the normality of each observation, and patients with a higher base-line mortality probability than the average patient will be assigned a higher weight.

The set method provides a graphical representation, called grass plot, to assist decision making. The drawbacks of this method lie in the complexity in determining the paired thresholds and interference based on the time between events rather than individual observations, which may cause delay in change detection.

(3) *Risk-adjusted CUSUM chart*: Cumulative Sum (CUSUM) control charts is a popular SPC technique due to their optimal properties. In the general setting, such charts aim to test the following hypothesis:

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

where θ denotes the parameter of the risk adjustment model, θ_0 is the base-line which is typically known, and θ_1 is a hypothesized value of interest. The following statistic is monitored

$$C_t = \max\ (0, C_{t-1} + \mathrm{W_t})$$

where W_t is the CUSUM score assigned to the t^{th} observation. A control limit H will be found through simulation to achieve a specified in-control average run length (ARL_0), and an alarm is signaled when $C_t > H$. The CUSUM score is given by the log-likelihood ratio of the two hypotheses

$$W_t = \log\left(\frac{\mathrm{L}(\theta_1|y_t)}{\mathrm{L}(\theta_0|y_t)}\right)$$

Where y_t is the t^{th} observation, and $\mathrm{L}(\theta|y_t)$ is the likelihood function of the risk adjustment model. For example, for binary data following a Bernoulli distribution with parameter θ=p, the likelihood function is

$$\mathrm{L}(\theta_1|y_t) = \theta^{\mathrm{yt}}\ (1 - \theta)^{\mathrm{1\text{-}yt}}$$

CUSUM charts based on the above likelihood have been used widely to monitor defective rate of products in industrial processes.

Risk-adjusted CUSUM charts for binary performance measure are first proposed by Steiner, et al. (2000) in monitoring 30-day mortality in cardiac surgeries, and then applied in other applications such as liver transplant to monitor one-year mortality (Leandro, Rolando, and Gallus, 2005) and coronary artery bypass surgeries to monitor adverse outcomes (Novick, et al, 2006). These charts, like their non-risk-adjusted counterparts in industrial contexts, are very powerful in detecting small changes in performance, but their use is limited by the perceived difficulty of interpretation by health care practitioners (Cook, Coory and Webster, 2011; Pilcher, et al, 2010).

(4) *Risk-adjusted EWMA chart*: Like CUSUM charts, the exponentially weighted moving average (EWMA) charts are a popular and widely used SPC technique. The statistic monitored in these charts takes the following form

$$C_t = \gamma S_t + (1-\gamma)C_{t-1}$$

Where S_t is the EWMA score assigned to the t^{th} observation, and $0 < \gamma \leq 1$ is a smoothing- constant. Essentially, the statistic is a linear combination of all the observations with higher weights assigned to recent observations. With the linearity in the statistic, its distribution can be obtained analytically, and consequently control limits can be specified base on that.

There are different definitions for the EWMA score depending on the types of data monitored. For binary performances, S_t can be the base-line mortality probability or the difference of the observed and the base line mortality probability (Cook, et al., 2008; Cook, Coory and Webster, 2011).

The risk-adjusted EWMA charts have similar performance to the risk- adjusted CUSUM charts in detecting small changes. It's main advantage over the latter lies in its intuitive interpretation as the EWMA statistic can be viewed as an estimate of the current level of the process. Moreover, the influence of previous observations is removed in the statistic gradually by adjusting the weights rather than resetting the statistic as CUSUM does.

Bayesian Approaches: Bayesian approaches have been used for process monitoring and change detection in various applications. Recently, such approaches are developed for different risk- adjusted monitoring problems, including Phase I monitoring (Assareh, Smith and Mengersen, 2011a, 2011b; Assareh and Mengersen, 2012), estimating the location where change in performance occurs (Assareh, Smith and Mengersen, 2011c), and self–starting performance monitoring (Zeng and Zhou, 2011). As suggested by Assareh, Smith and Mengersen (2011a, 2011c), Bayesian approaches can be used in conjunction with the non-Bayesian control charts such as risk-adjusted CUSUM charts to estimate the location of the change point when a change is detected using those charts. Summaries, such as mean, median and mode, of the posterior samples can be used as estimates of the change point. The drawbacks of Bayesian approaches lie in its need for specifying prior distributions and computation load.

4.1.2 Phase I Monitoring

The most popular method for phase I change detection is the generalized likelihood ratio (GLR) method due to its generality (Lai, 1995). It has been used in various applications such as profile data (Mahmoud et al., 2007; Kazemzadeh et al., 2008) and simple logistic models (Kamran et al., 2012). Two types of changes may take place in practice:

Change in model form: In case of a change in model form the data follow a model form such as linear, quadratic, cubic or any polynomial before the change-point and a completely different model form after the change-point. Figure 4.2 shows an example of change in model form. In this example the data from x=0 to 6 follows a linear equation, while the data from x=7 follows a fourth-order polynomial model.

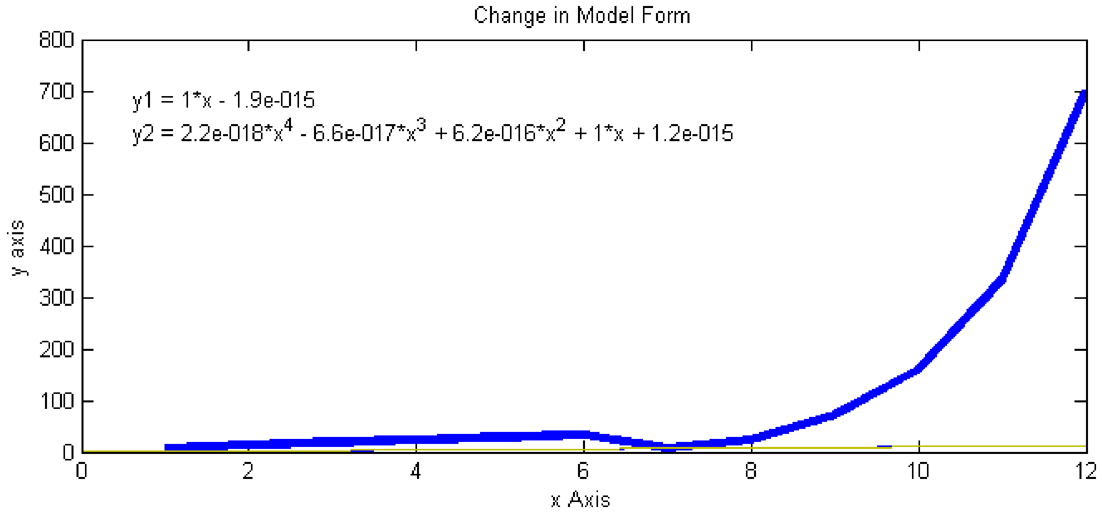


Figure 4-2 Change in Model Form

Change in model parameter**:** In case of a change in the model parameters there can be a change in one or more parameters of the model. For example, as shown in Figure 4.3 the data follows a linear model. However, before the change point the model follows the model of y1 and after the change point the model follows the model of y2. The model form however remains linear.

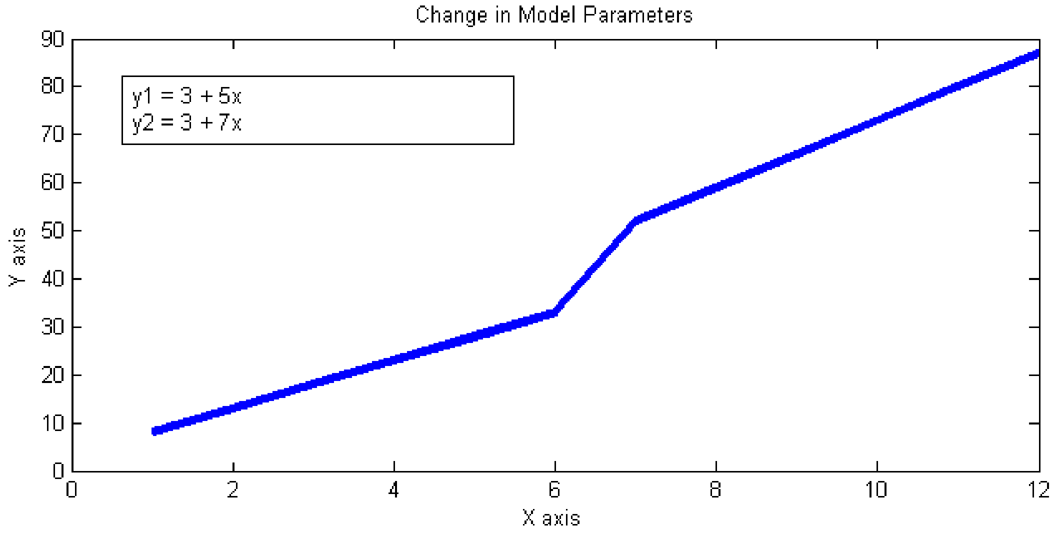


Figure 4-3 Change in Model Parameters

In this research we will focus on the second type of changes, i.e., changes in the model parameters. As a result, this method is built on the following change- point model

$$y_i \sim \begin{cases} M(y_i \mid \boldsymbol{\theta}_0) & i = 1,....,k \\ M(y_i \mid \boldsymbol{\theta}_1) & i = k+1,....,m \end{cases}$$

where $M(\bullet|\theta_0)$ is an appropriate statistical model of the data with parameters $\boldsymbol{\theta}$, *m* is the total number of available observations in the historical dataset, and *k*, 1≤*k*≤*m*, is the change point at which the model parameter changes from $\boldsymbol{\theta}_0$ to $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_0 \neq \boldsymbol{\theta}_1$. The change detection is equivalent to testing the hypothesis

$$H_0 : k = m$$

$$H_1 : 1 \leq k \leq m-1$$

When the null hypothesis is rejected, we conclude that change occurred in the process; otherwise we conclude that the process is in control.

Assume the observations are independent of each other. The likelihood under the null hypothesis is

$$L(y_1,......y_m) \mid H_0) = L(y_1,......y_m \left| \hat{\boldsymbol{\theta}}(y_1,......y_m)\right)$$

where $\hat{\boldsymbol{\theta}}(y_1,......y_m)$ is the estimate of $\boldsymbol{\theta}$ based on all the observations. The likelihood under the alternative hypothesis is

$$L(y_1,......y_m) \left|H_1\right) = L(y_1,......y_k \left|\hat{\boldsymbol{\theta}}_0(y_1,......y_k)\right) \bullet L(y_{k+1},......y_m \left|\hat{\boldsymbol{\theta}}_1(y_{k+1},......y_m)\right)$$

where $\hat{\boldsymbol{\theta}}_0(y_1,......y_k)$ is the estimate of $\boldsymbol{\theta}_0$ based on all the observations { $y_1,......y_k$} and $\hat{\boldsymbol{\theta}}_1(y_{k+1},......y_m)$ is the estimate of $\boldsymbol{\theta}_1$ based on all the observations { $y_{k+1},......y_m$}. The likelihood ratio which is a function of the change-point "*k*" is

$$R(k) = \frac{L(y_1,......y_m)\left|H_1\right)}{L(y_1,......y_m)\left|H_0\right)}$$

$$R(k) = \frac{L(y_1,......y_k \left|\hat{\boldsymbol{\theta}}_0(y_1,......y_k)\right) \bullet L(y_{k+1},......y_m \left|\hat{\boldsymbol{\theta}}_1(y_{k+1},......y_m)\right)}{L(y_1,......y_m \left|\hat{\boldsymbol{\theta}}(y_1,......y_m)\right)}$$

Usually the logarithm of this ratio is used for convenience

$$LgR(k) = \log\left(\frac{L(y_1,......y_m)\left|H_1\right)}{L(y_1,......y_m)\left|H_0\right)}\right)$$

$$LgR(k) = \log\left(\frac{L(y_1,......y_k \left|\hat{\boldsymbol{\theta}}_0(y_1,......y_k)\right) \bullet L(y_{k+1},......y_m \left|\hat{\boldsymbol{\theta}}_1(y_{k+1},......y_m)\right)}{L(y_1,......y_m \left|\hat{\boldsymbol{\theta}}(y_1,......y_m)\right)}\right)$$

$$LgR(k) = \log\left(L(y_1,......y_k \left|\hat{\boldsymbol{\theta}}_0(y_1,......y_k)\right)\right) + \log\left(L(y_{k+1},......y_m \left|\hat{\boldsymbol{\theta}}_1(y_{k+1},......y_m)\right)\right)$$

$$-\log\left(L(y_1,......y_m \left|\hat{\boldsymbol{\theta}}(y_1,......y_m)\right)\right)$$

Since this value depends on the value of "*k*", the largest likelihood ratio among all the possible values of *k* will be used as the statistic in change detection:

$$c = \max_{1 \le k \le m-1} LgR(k)$$

The upper control limit (UCL) of this statistic can be obtained through Monte Carlo simulation of null samples. Specifically, *m* observations following the null model with parameter $\hat{\boldsymbol{\theta}}(y_1,......y_m)$ are generated first and the statistic *c* is calculated for the sample. This is repeated for a number of times, which produces a set of *c* values. The 100(1-*α*)% percentile of these *c* values will be used as the control limit, where *α* is the specified Type I error rate. When a change is detected, the true change point *K* can be estimated by

$$\hat{K} = \arg\max_{1 \le k \le m-1} LgR(k)$$

Figure 4.4 illustrates how the value of *c* and *K* can be found.

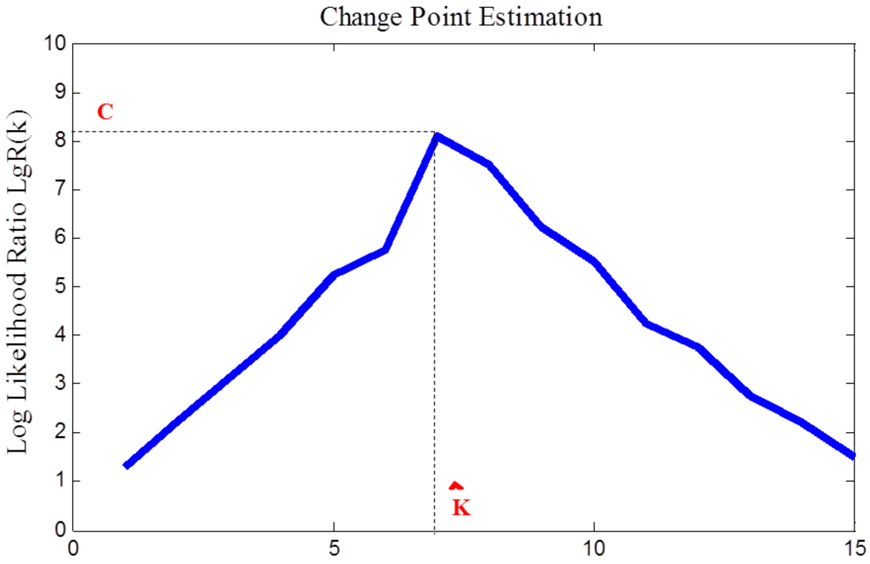


Figure 4-4 Change Point Estimation

In this study, we will apply the GLR method for Phase I monitoring of the readmission data.

4.2 Problem Formulation

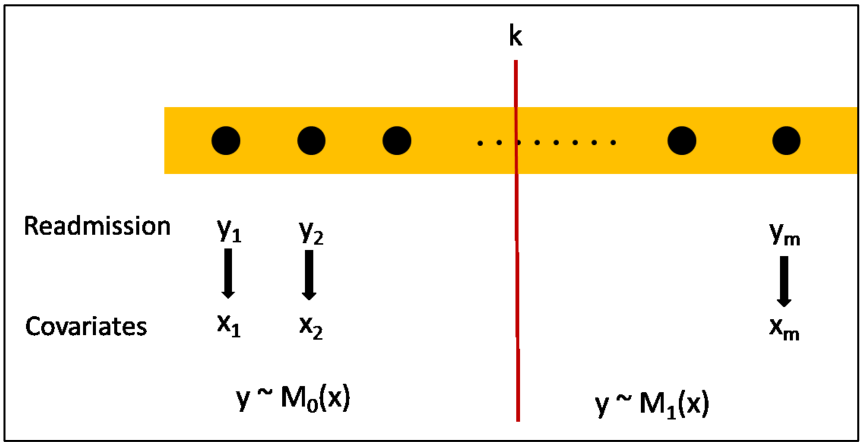


Figure 4-5 The Change Point Model Used in this Study

The change point model used in this study is shown in Figure 4.5. Assume there are totally *m* observations in the historical dataset and a change point *k* may exist in the data. The observations {(x_1, y_1),……, (x_k, y_k)} follow model M_0, while {(x_{k+1}, y_{k+1}),……, (x_m, y_m)} follow model M_1. Here the two models are of the same form, but with different parameters. According to the modeling study described in chapter 3, the model can be a logistic regression model or a logistic regression tree model. Both of these forms will be considered in this study. As in a typical Phase I analysis, here our problem is to determine if there is any change in the process, and if so, estimate the location of the change point. We focus on detecting a single change point in a given data set, and multiple change points can be detected by using the binary segmentation strategy described in Chapter 2.

There are two special issues in applying the GLR method for the readmission data:

Estimating the model parameters: To calculate the GLR statistic, we need to find estimates of the parameters under each model. For too small or too large *k* values, the observations used to estimate M_0 or M_1 will not be adequate to guarantee accurate estimation of the model parameters. This problem becomes more serious when a higher dimension of covariates is involved. Thus, appropriate *k* values need to be determined.

Determining the control limits: In applying Monte Carlo simulation to find the control limit, we need to simulate observations from the null hypothesis. This is easy in regular models without covariates. However our simulated data should mimic the real data and hence in the presence of covariates, this is complex as we need to simulate data of both the response and covariates. The simulated data should follow similar patterns as the observed data.

These two issues need to be addressed in this study.

4.3 The Proposed Approach

The proposed Phase I monitoring approach for readmission data has two critical components, the GLR statistic and the procedure to find the Upper Control Limit (UCL). Details of these two components are given as follows.

4.3.1 GLR statistic

To solve the issue (1) mentioned in Section 4.2, we will apply a window strategy for possible values of *k*, that is $k\epsilon[L_k, U_k]$, where L_k and U_k, $1< L_k < U_k < m$, are the lower and upper bound of the window, respectively. The two bounds can be specified according to the dimension of covariates in the model. For a higher dimension of covariates, a larger L_k and a smaller U_k should be used. Under this strategy, the GLR statistics for the GLR statistics for the LR model and the LRT model are derived in the following.

When the logistic regression model is used for the data, let **β** be the parameter vector of the model, and $X_i = [1\ x_i]$, where $1 = [1,\ldots..,1]'$ is a column vector of length *m*. The LR model is

$$\left.\begin{array}{l} y_i \sim Bin(p_i) \\ \log \dfrac{p_i}{1-p_i} = X_i\boldsymbol{\beta} \end{array}\right\} \Rightarrow P(y_i|\boldsymbol{\beta}) = \frac{e^{X_i\boldsymbol{\beta}y_i}}{1+e^{X_i\boldsymbol{\beta}}}$$

Derivation of the above is as follows:

$$\log \frac{p_i}{1-p_i} = X_i\boldsymbol{\beta}$$

$$\frac{p_i}{1-p_i} = e^{X_i\boldsymbol{\beta}}$$

$$p_i = e^{X_i\boldsymbol{\beta}}(1-p_i)$$

$$p_i = \frac{e^{X_i\boldsymbol{\beta}}}{1+e^{X_i\boldsymbol{\beta}}}$$

$$y_i \sim Bin\,(p_i) \Rightarrow y_i \sim Bin\left(\frac{e^{X_i\boldsymbol{\beta}}}{1+e^{X_i\boldsymbol{\beta}}}\right)$$

Since Binary variable can have only two values “0” and “1”,

$$\mathit{If}\ y_i = 1 \Rightarrow P\,(y_i|\,\boldsymbol{\beta}) = \left(\frac{e^{X_i\boldsymbol{\beta}}}{1+e^{X_i\boldsymbol{\beta}}}\right) \quad \textbf{OR} \quad P\,(y_i|\,\boldsymbol{\beta}) = \left(\frac{e^{X_i\boldsymbol{\beta}\,y_i}}{1+e^{X_i\boldsymbol{\beta}}}\right)$$

$$\mathit{If}\ y_i = 0 \Rightarrow P\,(y_i|\,\boldsymbol{\beta}) = 1 - \left(\frac{e^{X_i\boldsymbol{\beta}}}{1+e^{X_i\boldsymbol{\beta}}}\right) = \frac{1}{1+e^{X_i\boldsymbol{\beta}}} \quad \textbf{OR} \quad P\,(y_i|\,\boldsymbol{\beta}) = \left(\frac{e^{X_i\boldsymbol{\beta}\,y_i}}{1+e^{X_i\boldsymbol{\beta}}}\right)$$

Therefore,

$$\left.\begin{array}{l} y_i \sim Bin(p_i) \\ \log \dfrac{p_i}{1-p_i} = X_i\boldsymbol{\beta} \end{array}\right\} \Rightarrow P(y_i|\boldsymbol{\beta}) = \frac{e^{X_i\boldsymbol{\beta}y_i}}{1+e^{X_i\boldsymbol{\beta}}}$$

Accordingly, the change point model is

$$\begin{cases} P(y_i|\boldsymbol{\beta}) = \dfrac{e^{X_i\boldsymbol{\beta}_0\, y_i}}{1+e^{X_i\boldsymbol{\beta}_0}} & i \le k \\ P(y_i|\boldsymbol{\beta}) = \dfrac{e^{X_i\boldsymbol{\beta}_1 y_i}}{1+e^{X_i\boldsymbol{\beta}_1}} & i > k \end{cases}$$

Assuming that the events are independent of each other, the likelihood under the null model (i.e. no change) is

$$L\,(y_1,\ldots\ldots\ldots\ldots, y_m\big|\,\hat{\boldsymbol{\beta}}) = \prod_{\mathrm{i=1}}^{\mathrm{m}} \mathrm{P}(\mathrm{y_i}\big|\,\hat{\boldsymbol{\beta}}) \;=\; \prod_{\mathrm{i=1}}^{\mathrm{m}} \left(\frac{e^{X_i\hat{\boldsymbol{\beta}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}}}\right)$$

Assuming the events are independent, the likelihood under the change point model is

$$L\,(y_1,\ldots., y_m\big|\,\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, k) = \prod_{\mathrm{i=1}}^{\mathrm{k}} \mathrm{P}(\mathrm{y_i}\,\Big|\,\hat{\boldsymbol{\beta}}_0). \prod_{\mathrm{i}=k+1}^{\mathrm{m}} \mathrm{P}\,(\mathrm{y_i}\big|\,\hat{\boldsymbol{\beta}}_1) \;=\; \prod_{\mathrm{i=1}}^{\mathrm{k}} \left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_0.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_0}}\right) .\prod_{\mathrm{i}=k+1}^{\mathrm{m}} \left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_1.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_1}}\right)$$

The maximum likelihood estimates (MLE) of the parameters in each model will be used to calculate the above likelihoods, that is

$$\begin{aligned} &\hat{\boldsymbol{\beta}} = \mathrm{MLE\,of}\ \boldsymbol{\beta}\ \text{based on}\ \{(x_1, y_1),\ldots\ldots,(x_m, y_m)\} \\ &\hat{\boldsymbol{\beta}}_0 = \mathrm{MLE\,of}\ \boldsymbol{\beta}\ \text{based on}\ \{(x_1, y_1),\ldots\ldots,(x_k, y_k)\} \\ &\hat{\boldsymbol{\beta}}_1 = \mathrm{MLE\,of}\ \boldsymbol{\beta}\ \text{based on}\ \{(x_{k+1}, y_{k+1}),\ldots\ldots,(x_m, y_m)\} \end{aligned}$$

The log likelihood ratio is

$$LgR(k) = \log\left(\frac{L\,(y_1,\ldots., y_m\big|\,\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_1, k)}{L\,(y_1,\ldots\ldots\ldots\ldots, y_m\big|\,\hat{\boldsymbol{\beta}})}\right) = \log\left(\frac{\displaystyle\prod_{\mathrm{i=1}}^{\mathrm{k}} \left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_0.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_0}}\right) .\prod_{\mathrm{i}=k+1}^{\mathrm{m}} \left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_1.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_1}}\right)}{\displaystyle\prod_{\mathrm{i=1}}^{\mathrm{m}} \left(\frac{e^{X_i\hat{\boldsymbol{\beta}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}}}\right)}\right)$$

$$= \sum_{i=1}^{k} log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_0.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_0}}\right) + \sum_{i=k+1}^{m} log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_1.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_1}}\right) - \sum_{i=k+1}^{m} log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}}}\right)$$

And the GLR statistic is

$$c = \max_{1 \le k \le m-1} LgR(k)$$

$$\hat{K} = \arg\max_{1 \le k \le m-1} LgR(k)$$

When the logistic regression tree model is used for the data, the GLR statistic for two subspaces will be derived as an example. Let the LRT model be with two subspaces. As an example this is shown in the schematic representation in Figure 4.6.

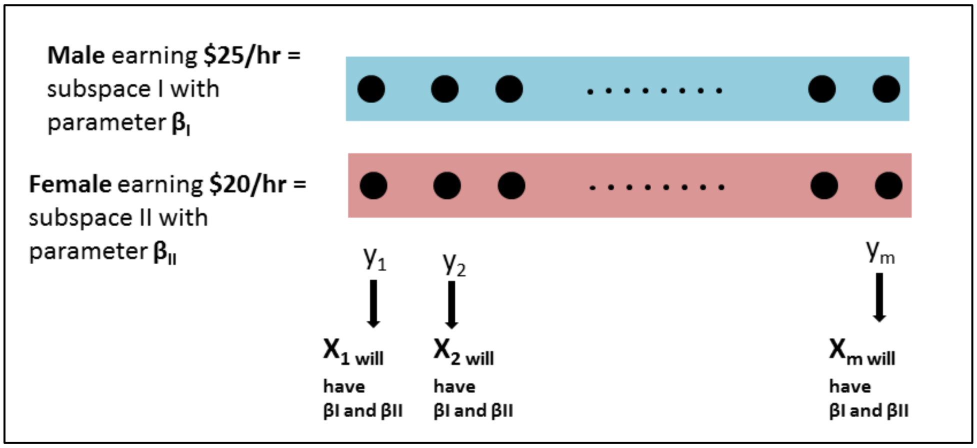

Male earning $25/hr = subspace I with parameter β_I
Female earning $20/hr = subspace II with parameter β_II
y1
y2
ym
X1 will have βI and βII
X2 will have βI and βII
Xm will have βI and βII

Figure 4-6 LRT Model Schematic Representations Under Null Hypothesis (No Change)

$$P\left(y_i \middle| \boldsymbol{\beta}_{\mathrm{I}}, \boldsymbol{\beta}_{\mathrm{II}}\right) = \begin{cases} \dfrac{e^{X_i \boldsymbol{\beta}_{\mathrm{I}} y_i}}{1+e^{X_i \boldsymbol{\beta}_{\mathrm{I}}}} & \mathrm{x_i} \in \mathrm{subspace\,I} \\ \dfrac{e^{X_i \boldsymbol{\beta}_{\mathrm{II}} y_i}}{1+e^{X_i \boldsymbol{\beta}_{II}}} & \mathrm{x_i} \in \mathrm{subspace\,II} \end{cases}$$

The schematic representation for the change – point model in LRT is shown in Figure 4.7.

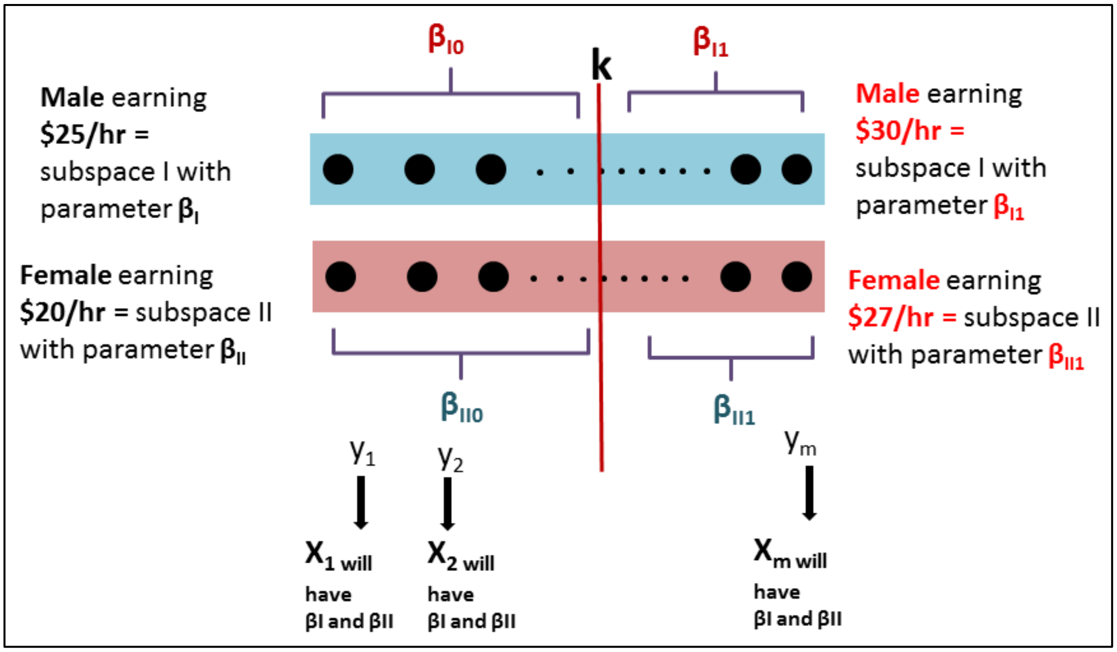


Figure 4-7 LRT Model Schematic Representation Under Alternate Hypothesis (Change Point *k*)

And the change-point model is

$$P\left(y_i \middle| \boldsymbol{\beta}_{\mathrm{I0}}, \boldsymbol{\beta}_{\mathrm{II0}}\right) = \begin{cases} \dfrac{e^{X_i \boldsymbol{\beta}_{\mathrm{I0}} y_i}}{1+e^{X_i \boldsymbol{\beta}_{\mathrm{I0}}}} & \mathrm{x_i} \in \text{subspace I} \\ \dfrac{e^{X_i \boldsymbol{\beta}_{\mathrm{II0}} y_i}}{1+e^{X_i \boldsymbol{\beta}_{\mathrm{II0}}}} & \mathrm{x_i} \in \text{subspace II} \end{cases} \qquad i \le k$$

$$P\left(y_i \middle| \boldsymbol{\beta}_{\mathrm{I1}}, \boldsymbol{\beta}_{\mathrm{II1}}\right) = \begin{cases} \dfrac{e^{X_i \boldsymbol{\beta}_{\mathrm{I1}} y_i}}{1+e^{X_i \boldsymbol{\beta}_{\mathrm{I1}}}} & \mathrm{x_i} \in \text{subspace I} \\ \dfrac{e^{X_i \boldsymbol{\beta}_{\mathrm{II1}} y_i}}{1+e^{X_i \boldsymbol{\beta}_{\mathrm{II1}}}} & \mathrm{x_i} \in \text{subspace II} \end{cases} \qquad i > k$$

The likelihood under the null model (i.e. no change) is

$$L\left(y_1,......, y_m \middle| \hat{\boldsymbol{\beta}}_{\mathrm{I}}, \hat{\boldsymbol{\beta}}_{\mathrm{II}}\right) = \prod_{\mathrm{i=1}}^{\mathrm{m}} \prod_{\text{subspaceI}} \left(\frac{e^{X_i \hat{\boldsymbol{\beta}}_{\mathrm{I}}.yi}}{1+e^{X_i \hat{\boldsymbol{\beta}}_{\mathrm{I}}}}\right) \cdot \prod_{\text{subspaceII}} \left(\frac{e^{X_i \hat{\boldsymbol{\beta}}_{\mathrm{II}}.yi}}{1+e^{X_i \hat{\boldsymbol{\beta}}_{\mathrm{II}}}}\right)$$

There is a double multiplication because we have assumed that the events as well as the subspaces are independent of each other. The likelihood under the change–point model is

$$L\,(y_1,\ldots,y_m|\,\hat{\boldsymbol{\beta}}_{\mathrm{I0}},\hat{\boldsymbol{\beta}}_{\mathrm{II0}},\hat{\boldsymbol{\beta}}_{\mathrm{I1}},\hat{\boldsymbol{\beta}}_{\mathrm{II1}},k)=\prod_{\mathrm{i=1}}^{\mathrm{k}}\prod_{\mathrm{subspaceI}}\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I0}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I0}}}}\right)\cdot\prod_{\mathrm{subspaceII}}\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II0}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II0}}}}\right)$$

$$\times\prod_{\mathrm{i}=k+1}^{\mathrm{m}}\prod_{\mathrm{subspaceI}}\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I1}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I1}}}}\right)\cdot\prod_{\mathrm{subspaceII}}\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II1}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II1}}}}\right)$$

Similarly, we can obtain the log-likelihood

$$LgR(k)=\log\left(\frac{L\,(y_1,\ldots.,y_m|\,\hat{\boldsymbol{\beta}}_{\mathrm{I0}},\hat{\boldsymbol{\beta}}_{\mathrm{II0}},\hat{\boldsymbol{\beta}}_{\mathrm{I1}},\hat{\boldsymbol{\beta}}_{\mathrm{II1}},k)}{L\,(y_1,\ldots\ldots\ldots.,y_m|\,\hat{\boldsymbol{\beta}}_{\mathrm{I}},\hat{\boldsymbol{\beta}}_{\mathrm{II}})}\right)$$

$$=\sum_{i=1}^{k}\left[\sum_{subspaceI}\log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I0}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I0}}}}\right)+\sum_{subspaceII}\log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II0}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II0}}}}\right)\right]$$

$$+\sum_{i=k+1}^{m}\left[\sum_{subspaceI}\log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I1}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I1}}}}\right)+\sum_{subspaceII}\log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II1}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II1}}}}\right)\right]$$

$$-\sum_{i=1}^{m}\left[\sum_{subspaceI}\log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{I}}}}\right)+\sum_{subspaceII}\log\left(\frac{e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II}}.yi}}{1+e^{X_i\hat{\boldsymbol{\beta}}_{\mathrm{II}}}}\right)\right]$$

And the GLR statistic is

$$c=\max_{1\le k\le m-1}LgR(k)$$

$$\hat{K}=\underset{1\le k\le m-1}{\arg\max}\,LgR(k)$$

4.3.2 Procedure to find UCL

To address issue (2) described in Section 4.2, bootstrap sampling techniques (Efron and Tibshirani, 1993; Phaladiganon et al., 2011) will be applied in the simulation for finding the UCL. The bootstrap method will be used to generate the values of the covariates from the observed data, thus making the simulated data follow the same pattern of the observed data. The procedure to find UCL is illustrated in Figure 4.8. It consists of four steps:

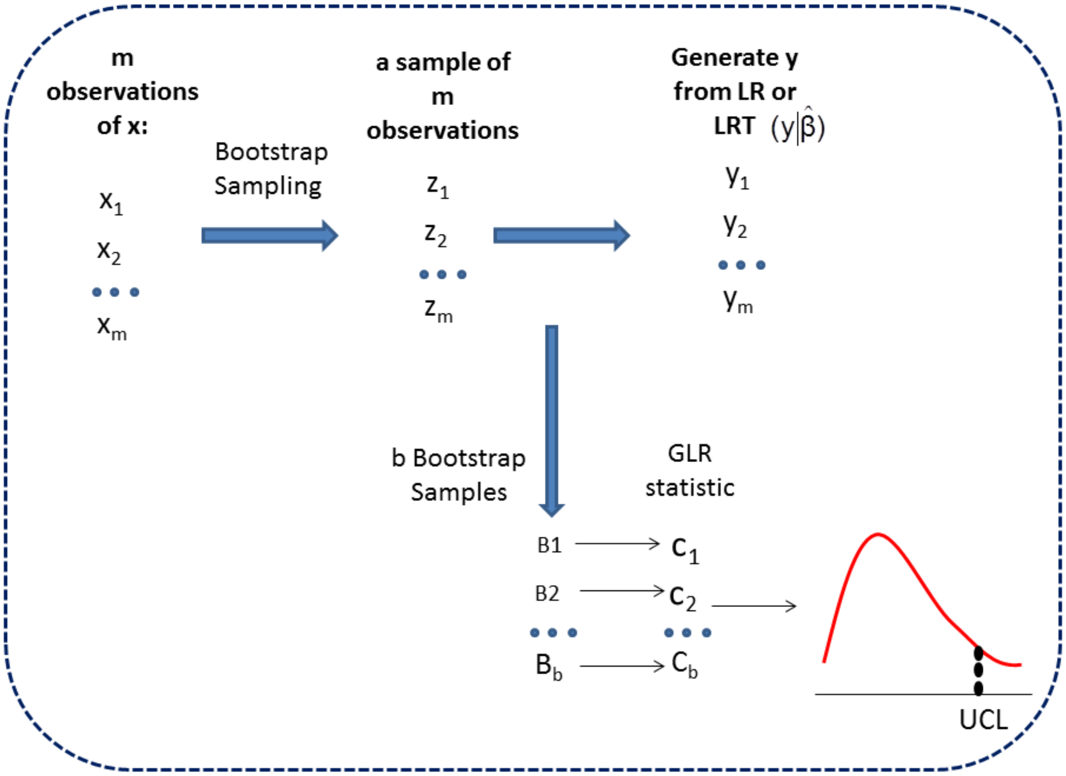


Figure 4-8 The Proposed Simulation Procedure to Find the Upper Control Limit

Step 1: *Generate sample of covariates*: Using bootstrap sampling techniques, a sample of the covariates involved in the model, $\{z_1,......,z_m\}$, is generated from the observations $\{x_1,......,x_m\}$. Note that we only need to simulate the values of the main effects of risk factors in the model. For covariates that are interactions of risk factors, their values can be obtained from the simulated values of the main effects. Also, when there is more than one covariate in the model, multivariate bootstrap sampling needs to be used.

Step 2: *Generate response values*. The response values corresponding to the simulated covariate values, $\{ y_1,......, y_m \}$, are generated based on LR or LRT model estimated using the whole historical dataset. The model should have been established in the modeling task described in Chapter 3.

Step 3: *Calculate GLR statistic*. Let B=$\{ z_1,......, z_m ; y_1,......, y_m \}$ be the bootstrap sample generated in Step 1 and 2. Calculate the GLR statistic "*c*" for this sample.

Step 4: *Obtain UCL*. Repeating the above steps for *b* times will lead to *b* values of the GLR statistic: $\{ c_1,......, c_b \}$. The UCL is the upper 100(1-α)% percentile of these values.

4.4 Simulation Study

Simulations have been done to validate the effectiveness of the proposed approach. Two base-line models are considered in this study: a LR model with a single covariate (Study 1) and a LRT with two covariates (Study 2). Under each model, a dataset without change and a dataset with change are simulated and the proposed approach is applied to the data for change detection. Parameter settings and the results of analysis are given as follows.

4.4.1 Study 1

Assume the data follow a LR model with a single covariate

$$\begin{aligned} & y_i \sim Bin\,(p_i) \\ & \log \frac{p_i}{1-p_i} = a + b\mathrm{x}_i \end{aligned}$$

This model has been used in many existing literature on surgical performance monitoring (e.g., Steiner et al., 2000) where the response represents the 30 day mortality (death/survival) of patients. Here the covariate is the patient risk score (e.g., parsonnet score), which measures the combined effect of patient risk factors. A dataset with *m*=500 observations, and x~uniform [0,71] are generated. Two cases are considered for the model parameter setting:

CASE I: The process is in control following the base line parameters a_0= -4, and b_0 = 0.07.

CASE II: The process changes at K=250, with a_0=-4, b_0=0.07 and a_1=-4, b_1=0.1.

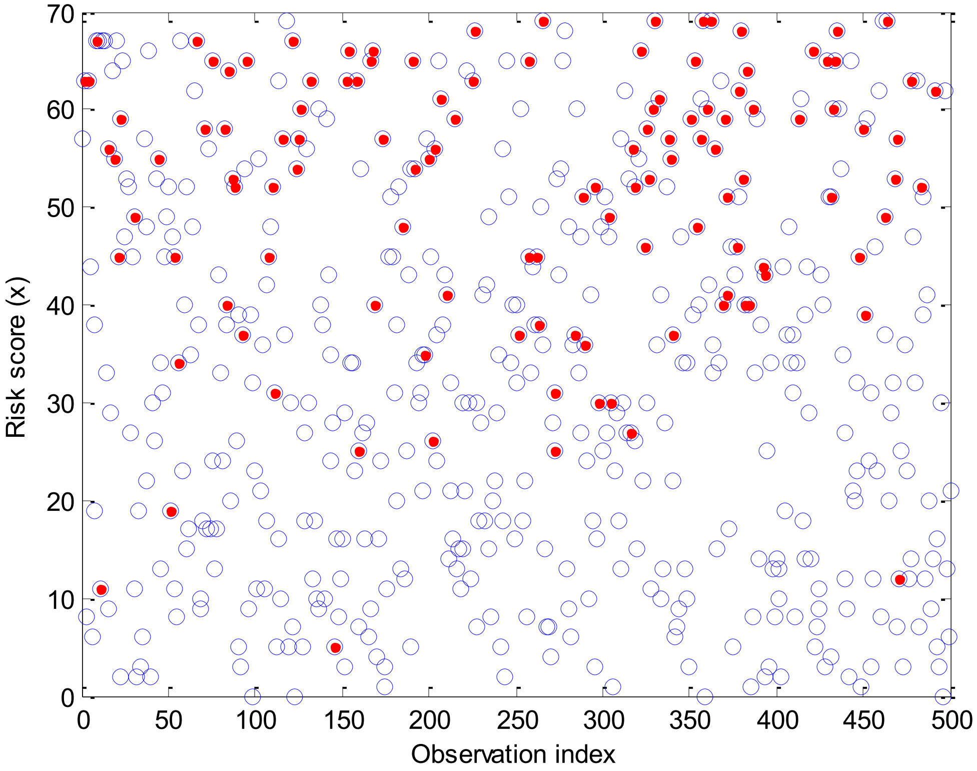


Figure 4-9 Simulated Data from the LR Model Without Change

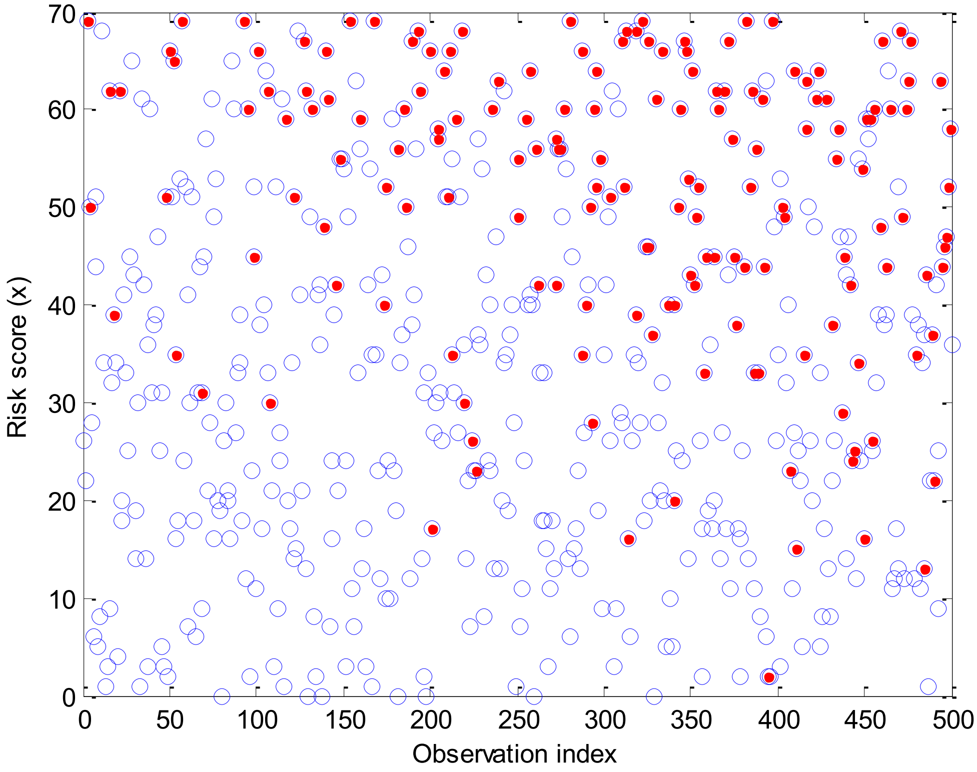


Figure 4-10 Simulated Data from the LR Model with Change (*K*=250)

Figure 4.9 and 4.10 show the simulated datasets under these two cases. The y axis in the figure denotes the risk score of each patient, and patients with y=1 are marked by solid dots. We can see that there seem no considerable change in Figure 4.9 as the distribution of solid dots is similar throughout the whole data set. In contrast, in Figure 4.10, the solid dots in the later segment of data have an apparently denser distribution than the earlier segment, indicating that there might be a significant change in the model parameters.

Results of Case I :To apply the proposed approach to the simulated dataset under CASE I, we first estimate the model parameters using the whole dataset. The parameter estimates are

$$\hat{a} = -3.8784, \hat{b} = 0.0661$$

In calculating the GLR statistic, the window boundaries are set to be L_k=50 and U_k=N−50=450 to ensure adequate samples for model estimation. In implementing the procedure in Section 4.3.2 to find the UCL, 5000 simulations are done. In each simulation, a sample of m=500 x values is generated from the observed x values, and then the corresponding y values are generated from the LR model with parameters being estimated values. The histogram of the calculated GLR statistics of the simulated samples is shown in Figure 4.11. Clearly, the distribution of GLR statistic is not normal, but right skewed. Given $\alpha = 0.05$, the UCL = 6.0157.

The log-likelihood ratio of the simulated dataset in Case I is calculated for each possible value of k. The results are shown in Figure 4.12. Note that no results exists for [1,50] and [451,500] due to the specified window. We can see that the likelihood ratios are all smaller than the UCL, meaning that the process is in control, which is consistent with the truth.

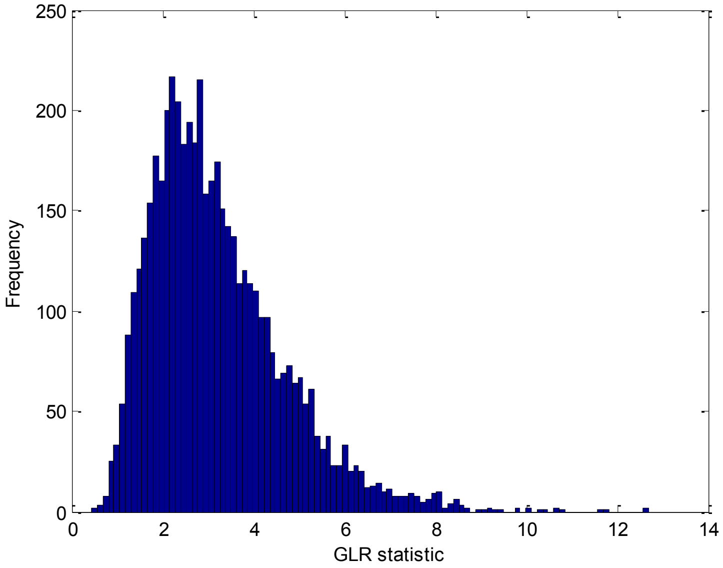


Figure 4-11 Histogram of GLR statistics in Case I of Study I

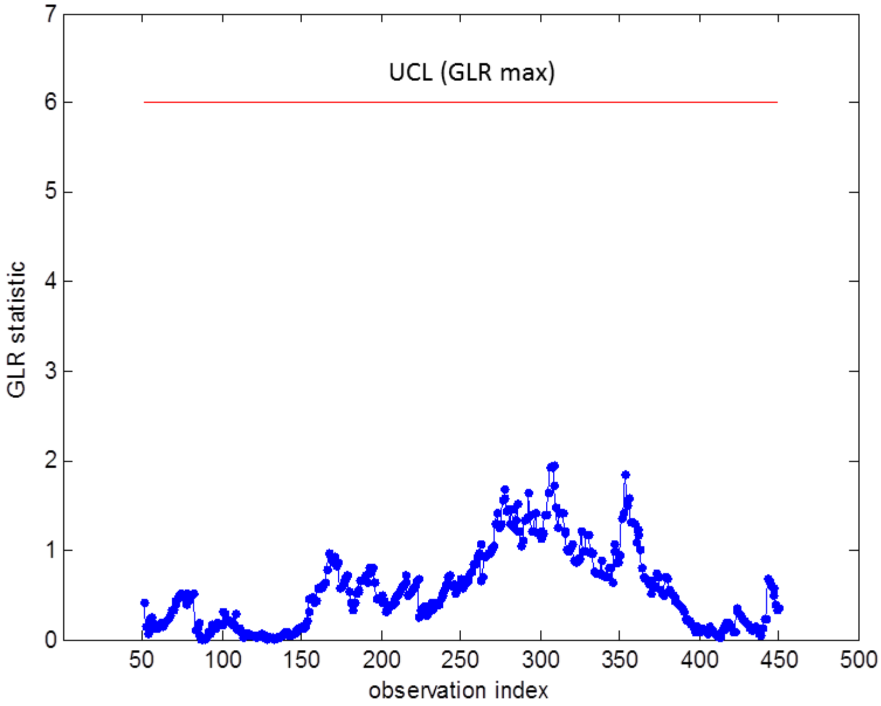


Figure 4-12 Log-likelihood Ratios of Simulated Dataset in Case I of Study I.

Results of Case II: The parameter estimates based on the whole dataset in Case II are

$$\hat{a} = -4.0833, \hat{b} = 0.0852$$

Like in Case I, 5000 simulations are done and the histogram of the resulting GLR statistics is shown in Figure 4.13. The UCL given $\alpha = 0.05$ is 6.1916, which is similar to the UCL in Case I. Figure 4.14 shows the log likelihood ratios of the simulated data for each possible value of *k*. We can see that a large portion of the likelihood ratios are beyond the UCL, indicating that there is a significant change in the data. The maximum value of the likelihood rations (i.e., the GLR statistic) is 22.3218, which is achieved when *k*=228. Thus, the estimate of the change point is 228, which is very close to the true change point *K*=250.

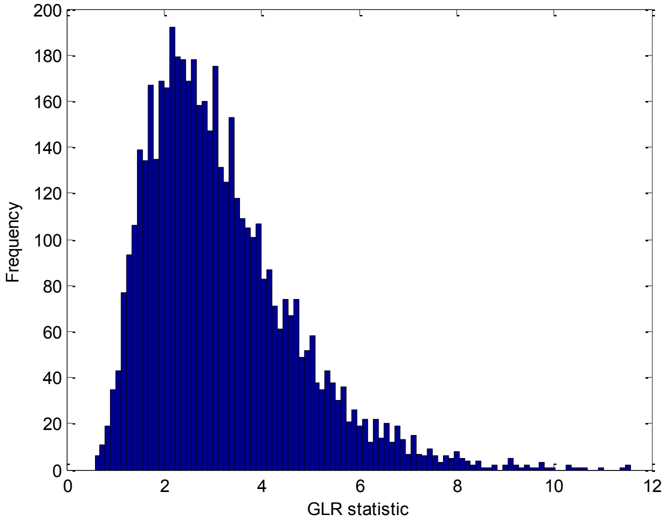


Figure 4-13 Histogram of the GLR Statistic in Case II of Study I

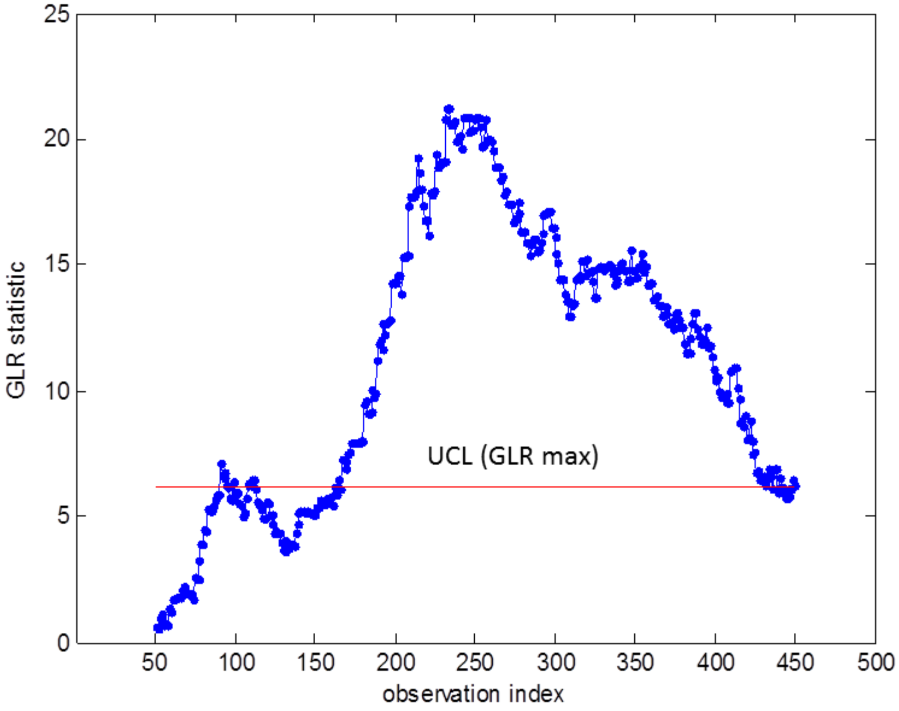


Figure 4-14 Log-likelihood Ratios of the Simulated Dataset in Case II of Study I

Distribution of change-point estimates: One important performance measure of the proposed approach is accuracy of change point estimates. Simulation has been done to obtain the distribution of change point estimates in Case II. Specifically, a dataset under Case II is generated and the proposed approach is applied to the data for change detection. If the change is detected, the change point estimates are obtained and saved. By repeating this 5000 times, a set of estimates are obtained. Figure 4.15 shows this distribution. Clearly, the change point estimates center at the true change point *K* =250, which suggest that when there are adequate observations, the change point can be accurately identified.

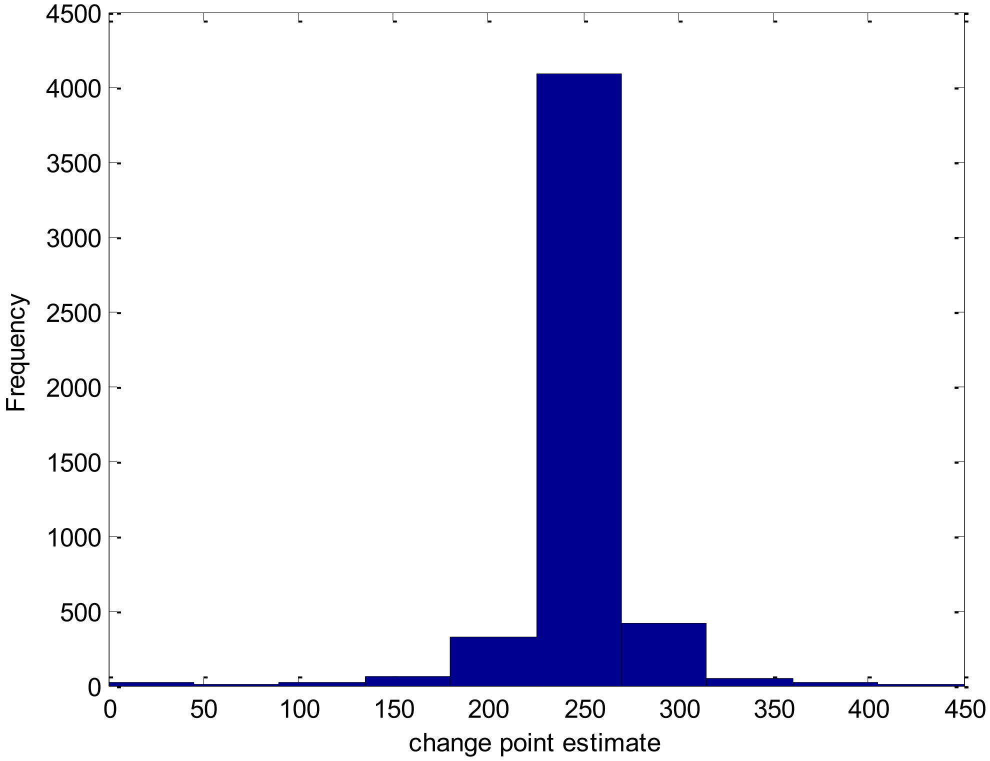


Figure 4-15 Distribution of Change Point Estimates with 5000 Simulations in Case II of Study I

4.4.2 Study 2

Assume the data follows a logistic regression tree model with two covariates, one of which is the patient risk score, and the other is the gender of patient (male=0, female =1). The male patients and female patients show different patterns in their surgical outcomes, i.e., the covariate “gender” is a splitting variable. Assume male/female patients account for 60% and 40% of the patient population, and their risk scores follow uniform [0,71] as in Study 1. Let the model for the two subspaces be

$$\text{Subspace I (male): } \begin{array}{l} y_i \sim Bin\,(p_i) \\ \log \dfrac{p_i}{1-p_i} = a_{\mathrm{I}} + b_{\mathrm{I}} \mathrm{x}_i \end{array}$$

$$\text{Subspace II (female): } \begin{array}{l} y_i \sim Bin\,(p_i) \\ \log \dfrac{p_i}{1-p_i} = a_{\mathrm{II}} + b_{\mathrm{II}} \mathrm{x}_i \end{array}$$

Two cases of the parameter setting are considered:

Case I: The process is in control following the base line parameters a_{10}=-4, b_{10}=0.07, $a_{\mathrm{II}0}$=-3.6, $b_{\mathrm{II}0}$=0.06.

Case II: The process changes at K=250, with $a_{\mathrm{I}0}$=-4, $b_{\mathrm{I}0}$=0.07, $a_{\mathrm{II}0}$=-3.6, $b_{\mathrm{II}0}$=0.06 and $a_{\mathrm{I}1}$=-4, $b_{\mathrm{I}1}$=0.1, $a_{\mathrm{II}1}$=-3.6, $b_{\mathrm{II}1}$=0.06.

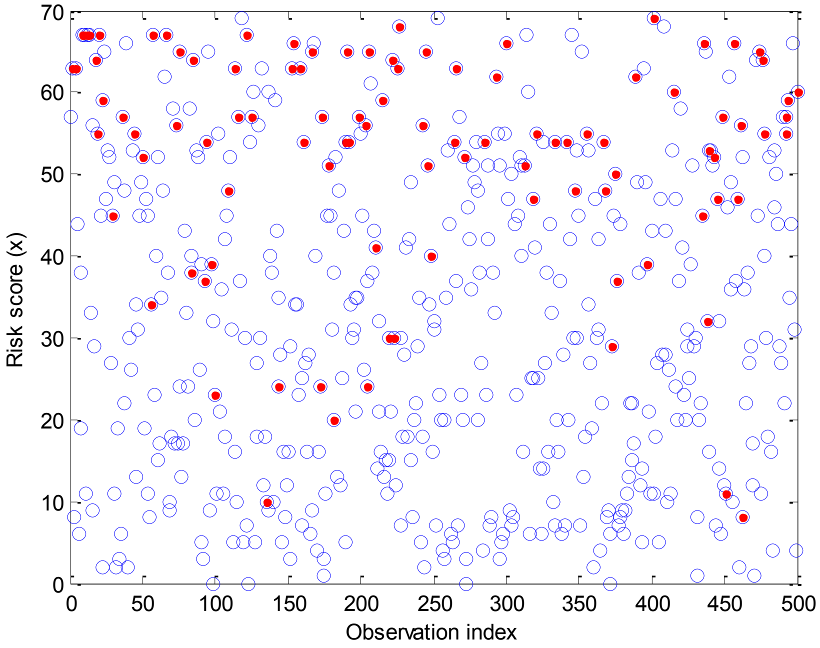


Figure 4-16 Simulated Data from the LRT Model Without Change

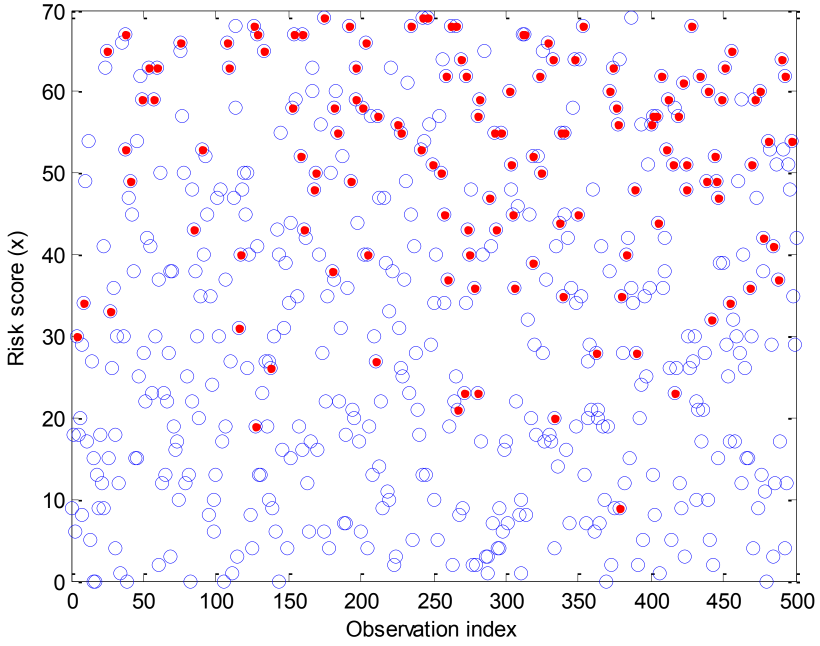


Figure 4-17 Simulated Data from the LRT Model with Change (*K*=250)

Figure 4.16 and Figure 4.17 show the simulated datasets under the two cases. The data in Case I show no evidence of change. The data in Case II show some evidence of change between the earlier segment and the later segment. Note that the change here is the same as that in Case II of study 1 (i.e., b_0=0.07 vs b_1=0.1). However, the evidence of change in Figure 4.17 is weaker than in Figure 4.4. This is due to the tree structure in Study 2 in which the change only happens to the model of male patients. Considering the complexity of model structure in Study 2, a wider window L_k=80 and U_k=420 is applied in calculating the GLR statistics.

Results of Case I

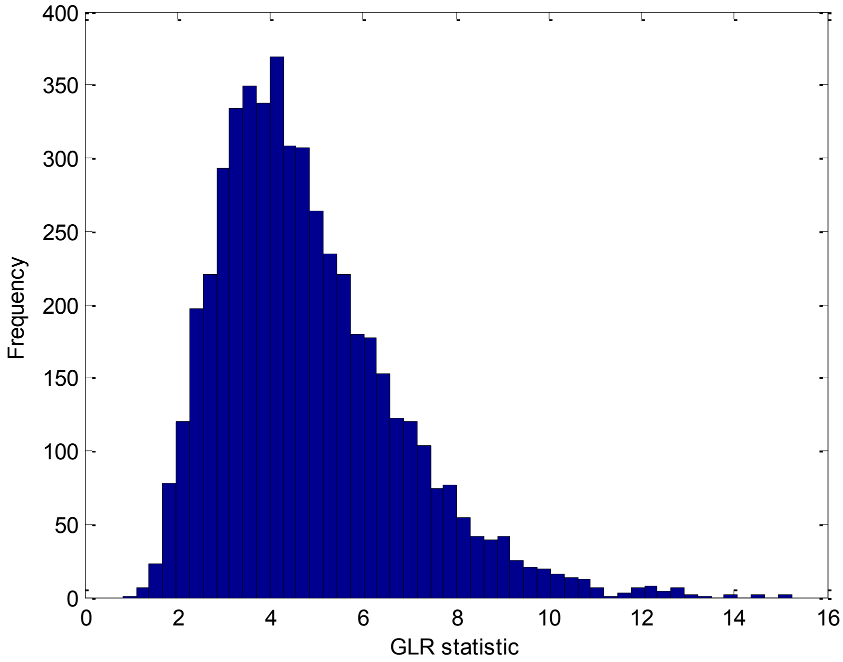


Figure 4-18 Figure 4.18 Histogram of the GLR Statistic in Case I of the Study 2

First, the LRT model is estimated using the whole dataset. The parameter estimates are

$$\hat{a} = -3.8784, \hat{b} = 0.0661$$

Then the procedure in Section 4.3.2 is applied to find the control limit using the above estimates. Figure 4.18 shows the histogram of the GLR statistics obtained in 5000 simulations. Given α=0.05, the UCL=8.5143. Finally, change detection is conducted to the simulated dataset. The log-likelihood ratios for different possible values of *K* are shown in Figure 4.19. Clearly, all the statistics are below the control limit, indicating that there is no change in the process.

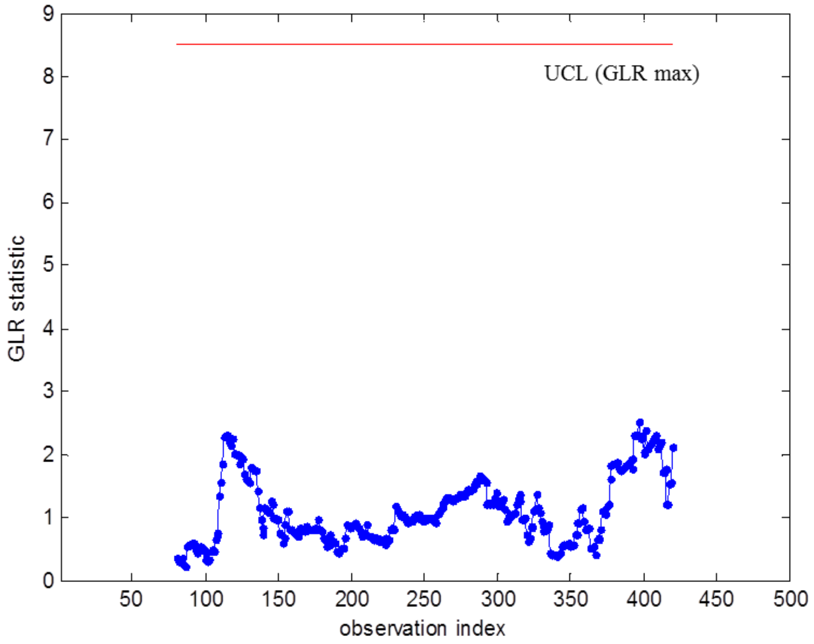


Figure 4-19 Log-likelihood Ratios of the Simulated Dataset in Case I of Study 2

Results of Case II

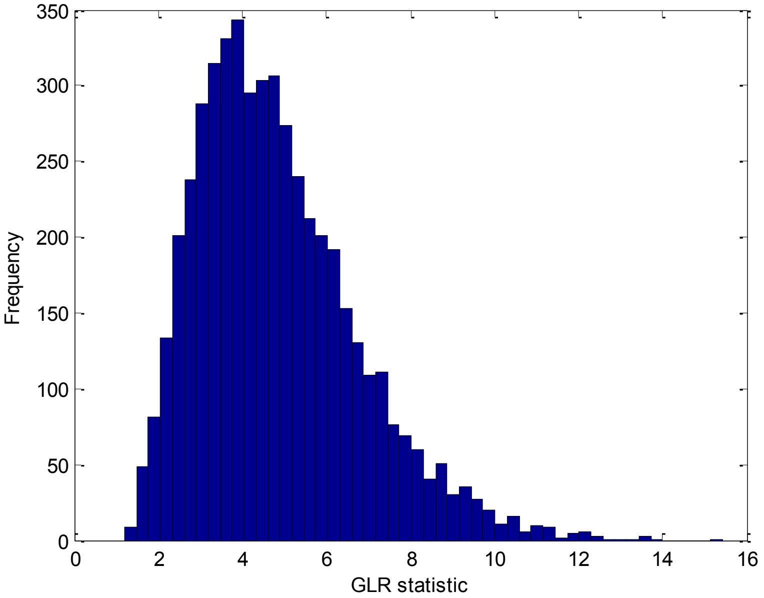


Figure 4-20 Histograms of the GLR statistic in Case II of Study 2

The parameter estimates of the LRT model based on the whole dataset are

$$\hat{a} = -3.8784, \hat{b} = 0.0661$$

which are used in the procedure to find the control limit. Figure 4.20 shows the GLR statistics obtained in 5000 simulations. Given α=0.05, the UCL=8.514. Figure 4.21 shows the log likelihood ratios for different possible values of *k*. The evidence of change is very strong based on the result in the figure. The highest point is 12.8561 and achieved when *k*=250, which is exactly the true change point. Again, note that the evidence of change in Figure 4.20 is weaker than in Figure 4.14 due to the tree structure of the model in Study 2.

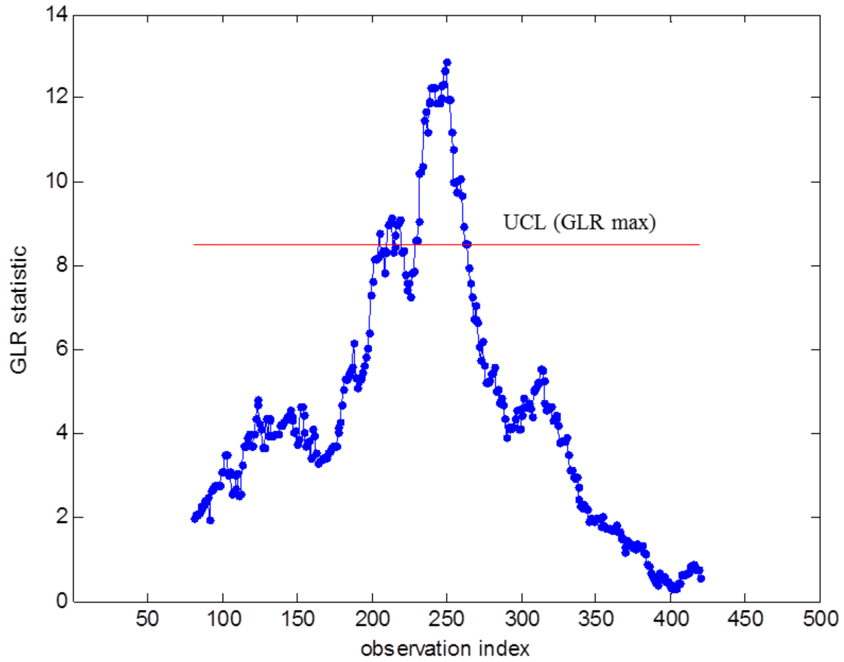


Figure 4-21 Log-likelihood Ratios of the Simulated Dataset in Case II of Study 2

Distribution of change point estimates: 5000 simulations are done to evaluate the performance of the proposed approach in change point estimation. In each simulation, a dataset following Case II is generated and change detection is concluded on the data. Figure 4.22 shows the histogram of the resulting change point estimates. We can see that just like in Study 1, the change point estimates center at the true change point *K*=250. However, the variance of these estimates seems to be larger than in Study 1, which, again, shows the effect of the tree structure.

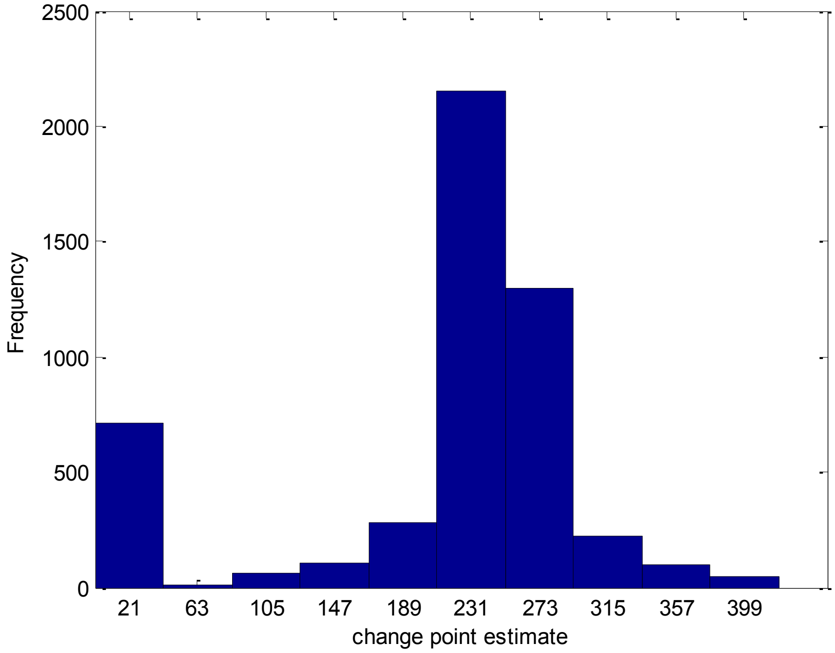


Figure 4-22 Distribution of Change Point Estimates in Case II of Study 2

4.5 Case Study

The proposed approach has been applied to the readmission dataset described in chapter 3 for phase I monitoring. The two models established in Chapter 3 are considered

LR Model

Logit (30-DAY READMISSION PROBABILITY) = -4.49 + 1.15 × LAMA + 1.96 × ER _VISITS + 0.19×HOSPITALIZATIONS + 0.33 × HOSPITALIZATIONS_COPD + 1.05× OXYGEN + 0.76×ALCOHOL_USE + 0.05× DRUG_USE – 0.91×ER_VISITS*DRUG_USE – 0.57×HOSPITALIZATIONS_COPD*OXYGEN

LRT Model

When OXYGEN = 0

Logit (30-DAY READMISSION PROBABILITY) = -3.82 + 1.93 × ER_VISITS + 0.63 × HOSPITALIZATIONS_COPD

When OXYGEN = 1 and DRUG_USE =0

Logit = (30-DAY READMISSION PROBABILITY) = -3.17 + 1.81× ER_VISITS + 0.33×HOSPITALIZATIONS – 0.33×HOSPITALIZATIONS_COPD

When OXYGEN=1 and DRUG_USE = 1

Logit (30-DAY READMISSION PROBABILITY) = -14.67 + 0.94×RDW

Since there are multiple covariates, we use a wider window than in the simulation study: L_k = 100 and U_k = 183. This means that if the true change point occurs<100 or >183, it cannot be identified. Change detection under each of the above models is conducted, and the results are reported in the following.

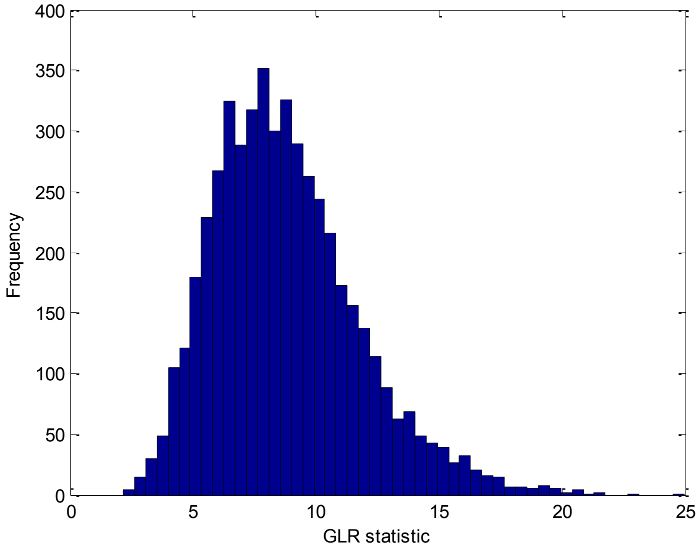


Figure 4-23 Histogram of the GLR statistic in Case Study under the LR model

We first find the control limit under the LR model. Figure 4.23 shows the histogram of the GLR statistics. Given α=0.05, UCL= 14.2822. Figure 4.24 shows the log likelihood ratios for different possible values of *k*. We can see that all these ratios are below the control limit, which indicates that there is no change in the care provider's performance during the considered period. Similar things are done under the under LRT model. Figure 4.25 shows the histogram in the simulation to find UCL and Figure 26 shows the GLR statistic for all possible *k*. The same conclusion is drawn, that is, there is no change during the considered period.

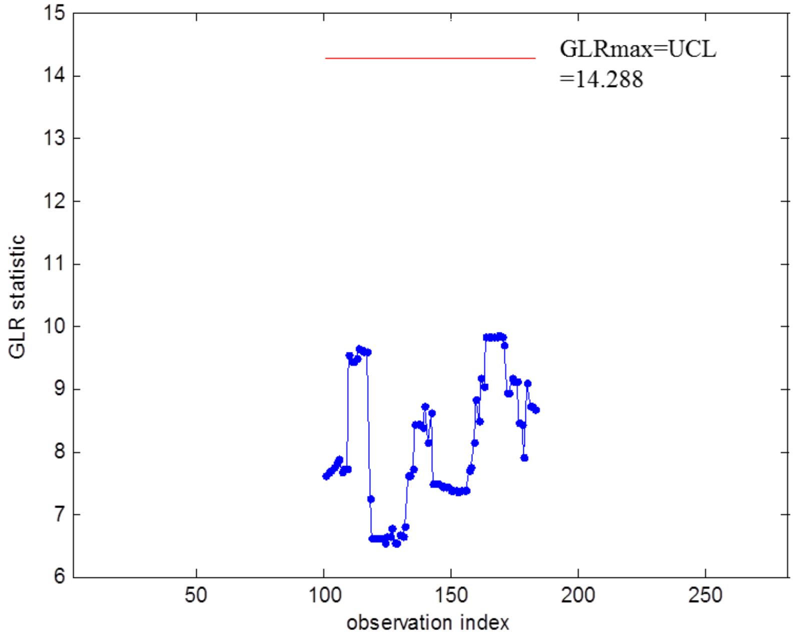


Figure 4-24 Log-likelihood Ratios of the Data Under the LR Model

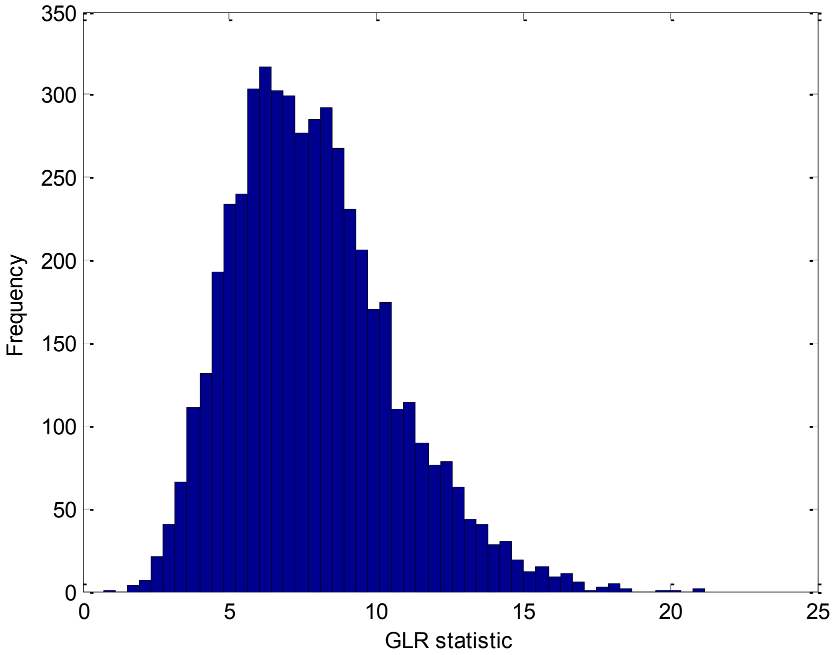


Figure 4-25 Histogram of the GLR statistics in Case Study under the LRT model

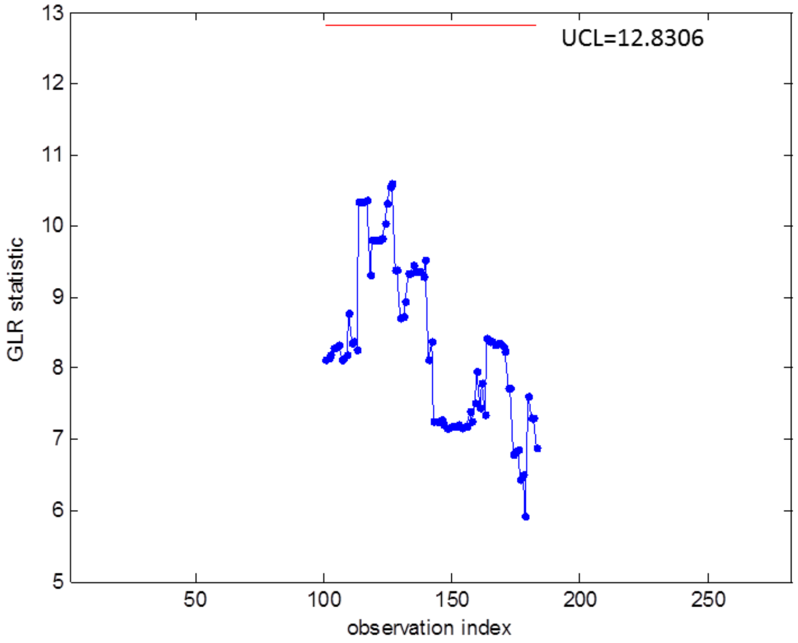


Figure 4-26 Log-likelihood Ratios of the Data Under the LRT model

Chapter 5

Summary and Future Work

In this research quality improvement and process control methods are developed to monitor complex systems. Two application areas are considered, complex manufacturing processes and healthcare delivery processes. Quality monitoring typically consists of two phases called Phase I analysis (or offline monitoring) and Phase II analysis (or online monitoring) (Sullivan, 2002). This research has focused on Phase I quality monitoring. Our findings are and future work are summarized as follows.

5.1 Quality Monitoring of Optical Profiles in Low-E Glass Manufacturing

The quality data generated in Low-E glass manufacturing are optical profiles which are one type of profile data. A profile, or a curve, represents the relationship of a response variable on an explanatory variable such as time and distance. This study is focused on the Phase I monitoring of optical profiles. A robust Phase I monitoring strategy for the optical profile data is developed. The proposed strategy has three steps. The first step is to fit an appropriate statistical model to the data to obtain estimates of the coefficients. The second step is to transform correlated multivariate coefficient estimates into univariate independent components using Independent Component Analysis (ICA), and the third step is to monitor selected Independent Components (IC) using univariate nonparametric control charts. We considered two methods to detect multiple change points: binary segmentation and sequential segmentation. Two numerical studies are done to understand the performance of this proposed strategy. In the first simulation study, performance of the two data segmentation methods for multiple change point detection is studied. In the second simulation study, properties of the proposed strategy for phase-I monitoring of profile data is studied. Finally, the proposed strategy was applied to the real world data obtained from a glass manufacturing company. Using this strategy, two groups of profiles were identified. The root causes of these changes were identified and investigated.

5.2 Risk-adjusted Readmission Monitoring in COPD Care

There are 3 components in this research: hospital readmission, chronic obstructive pulmonary diseases (COPD), and risk-adjustment. The existing quality monitoring work in healthcare is limited in the following aspects:

Firstly, they mostly focus on patient mortality in surgical/intensive unit care and no work has been done on the monitoring of patient readmission in chronic disease applications like COPD care. Secondly, the existing work only considers one covariate in the monitoring. Finally, most existing work on risk-adjusted monitoring focuses on Phase II analysis, and there is little work on Phase I analysis. In this research a systematic approach for Phase I monitoring of patient readmission is proposed. It consists of two tasks: building an appropriate statistical model and monitoring the change detection based on the established model in the first task. In the first task, two types of models were studied, Logistic regression model and the logistic regression tree model. Using cross validation technique the best model was found. In this case Logistic regression tree model worked better than the logistic regression model in predicting the 30 day readmission. Once the best model was found then control charts based on Generalized Likelihood Ratio method were applied to monitor the historical data. Two numerical simulation studies were conducted to understand the performance of the proposed strategy. In the first study the proposed strategy was applied to a simulated logistic regression model and change was detected. In the second study, the proposed strategy was applied to a simulated logistic regression tree model and the change was detected. Finally, this method was applied to the real data obtained from the UTMB for 30 day readmission of COPD patients. It is found that no change occurred during the considered period.

5.3 Future Work

This research will be continued in the future in two directions: First, it is very difficult to obtain health care data. UTMB was gathering the information regarding COPD patients for more than 2 years. This research started with 400 data points but, after cleaning the data, only 283 data points were available. This is not enough data for reliable detection of changes in patient readmission. In the case study, since there are multiple covariates, we used a very wider window (L_k = 100 and U_k = 183). That means only changes occurring during this period can be detected. The proposed approach will be done when larger dataset become available to generate more reliable findings. Second, although the proposed approach is demonstrated using readmission data in COPD care in this research, it actually has broad applicability across various healthcare applications. For example, the Phase I risk-adjusted monitoring can also be applied to patient mortality data in surgical care. New applications of the proposed approach will be considered in our future research.

References

(2011b). “Change Point Detection in Risk Adjusted Control Charts”, Statistical Methods for Medical Research, in press.

(2011c). “Bayesian Estimation of the Time of A Linear Trend in Risk-Adjusted Control Charts”, International Journal of Computer Science, 38(4):409-417.

A.C..-W.Lau,L.Y.-C Yam and E.Poon, Hospital readmission in patients with acute exacerbation of chronic obstructive pulmonary disease, Respiratory Medicine, Vol.95 (2001) 876-884.

Almagro, P., Barreiro, B., de Echagüen, A. O., Quintana, S., Carballeira, M. R., Heredia, J. L., and Garau, J., Risk factors for hospital readmission in patients with chronic obstructive pulmonary disease, Respiration, vol. 73, pp. 311-17, 2006.

Amiri, A., Jensen, W.A., and Kazemzadeh, R.B. (2009) A case study on monitoring polynomial profiles in the automotive industry. Quality and Reliability Engineering International, 26:509-520.

An Introduction to risk assessment and risk adjustment models, American Medical Association, 2009 http://www.ama-assn.org/resources/doc/psa/risk-assessment.pdf

Arasteh, D., Carmody, J., Lee, E.S., and Selkowitz, S. (2004) Window systems for high-performance buildings, W. W. Norton & Company, New York, NY.

Assareh, H., and Mengersen, K. (2012). “Change Point Estimation in Monitoring Survival Time”, PLoS One, 7(3):1:10.

Assareh, H., Smith, I., and Mengersen, K. (2011a). “Bayesian Change Point Detection in Monitoring Cardiac Surgery Outcomes”, Q Manage Health Care, 20(3):207-222.

Azzalini, A. (1985) A class of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12:171-178.

Azzalini, A., and Genton, M.G. (2008) Robust likelihood methods based on the skew-t related distributions. International Statistical Review, 76(1):106-129.

Bahadori, K., and FitzGerald, J. M., Risk factors of hospitalization and readmission of patients with COPD exacerbation − systematic review, International Journal of COPD, vol. 2, no. 3, pp. 241-51, 2007.

Carmody, J., Selkowitz, S., and Heschhong, L. (1996) Residential windows: a guide to new technologies and energy performance. W. W. Norton & Company, New York, NY.

Chan, K.-Y., and Loh, W.-Y., LOTUS: An algorithm for building accurate and comprehensible logistic regression trees, Journal of Computational and Graphical Statistics, vol. 13, pp. 826-52, 2004.

Cocking, J. GL., Cook, D. A., and Iqbal, R. K. (2006). “Process Monitoring in Intensive Care with the Use of Cumulative Expected Minus Observed Mortality and Risk-adjusted p Charts”, Critical Care, 10(1):R28.

Cook, D. A., Coory, M., and Webster, R. A. (2011). “Exponentially Weighted Moving Average Charts to Compare Observed and Expected Values for Monitoring Risk-adjusted Hospital Indicators”, BMJ Qual Saf, 20:469-474.

Cook, D. A., Duke, G., Hart, G. K., Pilcher, D., and Mullany, D. (2008). “Review of the Application of Risk-adjusted Charts to Analyse Mortality Outcomes in Critical Care”, Critical Care and Resuscitation, 10(3):239-251.

Cook, D. A., Stefan, S. H., Cook, R. J., Farewell, V. T., and Morton, A. P. (2003). “Monitoring the Evolutionary Process of Quality: Risk-adjusted Charting to Track Outcomes in Intensive Care”, Crit Care Med, 31(6):1676-82.

COPD fact sheet.http;// http://www.lung.org/lung-disease/copd/resources/facts-figures/COPD-Fact-Sheet.html , American Lung Association, February 2011 (accessed in May 2013).

David M Smith, Anita Giobbie-Hurder, Morris Weinberger, Eugene Z. Oddone, William G. Henderson, David A. Asch, Carol M. Ashton, John R. Feussner, Paulette Ginier, James M. Huey, Denise M. Hynes, Lawrence Loo and Charles E. Mengel, Predicting non-elective hospital readmissions: A multi-site study, Journal of Clinical Epidemiology 53 (2000) 1113-1118.

Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman and Sunil Kriplani, Risk Prediction Models for hospital readmission, JAMA October 19, 2011- Vol 306, No.15.

Ding, Y., Zeng, L., and Zhou, S. (2006) Phase I analysis for monitoring nonlinear profiles in manufacturing processes. Journal of Quality Technology, 38(3):199-216.

Efron, B., and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman & Hall/CRC.

Frost, K., Arasteh, D., and Eto, J. (1993) Savings for energy efficient windows: current and future savings from new fenestration technologies in the residential market. Lawrence Berkeley Laboratory Report 33965, Lawrence Berkeley Laboratory.

G.Gudmundsson, T Gislason, C. Janson, E. Lindberg, R. Hallin, C.S. Ulrik, E. Brondum, M.M. Nieminen, T. Aine and P.Bakke, Risk Factors for re-hospitalization in COPD role of health status, anxiety and depression, European respiratory journal 2005, 26: 414-419.

Gordon, L., and Pollak, M. (1994) An efficient sequential nonparametric scheme for detecting a change of distribution. The Annals of Statistics, 22:763-804.

Grigg, O., and Farewell, V. (2004). “An Overview of Risk-adjusted Charts”, J. R. Statist. Soc. A, 167:523-539.

Grigg, O., and Farewell, V. T. (2004). “A Risk-adjusted Sets Method for Monitoring Adverse Medical Outcomes”, Statistics in Medicine, 23:1593-1602.

Hackl, P., and Ledolter, J. (1991) A control chart based on ranks. Journal of Quality Technology, 23:117-124.

Hawkins, D.M., and Deng, Q. (2010) A nonparametric change-point control chart. Journal of Quality Technology, 42(2):165-173.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R., and Pentz, M. A., Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes, American Journal of Epidemiology, vol. 147, no. 7, pp. 694-703, 1998.

Hyvarinen, A., Harhunen, J., and Oja, E. (2001) Independent component analysis. Wiley, New York, NY.

J. Garcia-Aymerich, E Farrero, M A Felez, J Izquierdo, R M Marrades and J M Anto , Risk Factors of readmission to hospital for a COPD exacerbation: a prospective study, Thorax 2003, 58: 100-105.

Jensen, W. A., Birch, J. B., and Woodall, W. H. (2008) Monitoring Correlation within Linear Profiles Using Mixed Models. Journal of Quality Technology, 40(2):167-183.

Jensen, W.A. and Birch, J.B. (2009) Profile monitoring via nonlinear mixed models. Journal of Quality Technology, 41(1):18–34.

Jones, L., and Woodall, W. (1998) The performance of Bootstrap control charts. Journal of Quality Technology, 30:362-375.

JudithGarcia-Aymerich,CArme Hernandez, Albert Alonso, Alejandro Casas, Robert Rodriguez-Roshin, Joseph M. Anto and Josep Roca. Effects of an integrated care intervention on risk factors of COPD readmission, Elsevier Respiratory Machine (2007) 101, 1462-1469.

Kamran, P., Jin, J., and Yeh, A. B. (2012). “Phase I Risk-adjusted Control Charts for Monitoring Surgical Performance by Considering Categorical Covariates”, Journal of Quality Technology, 44(1):39-53.

Katayoon Bahadori and J mark Fitz Gerald, Risk factors of hospitalization and readmission of patients with COPD exacerbation – systematic review, International journal of COPD 2007:2 (3) 241-251.

Kathryn E. Martin, Deborah L. Rogal, and Sharon B. Arnold, Health Based rsik assessment: Risk adjusted payments and beyond, Academy health, January 20

Kazemzadeh, R.B., Noorossana, R., and Amiri, A. (2008) Phase I monitoring of polynomial profiles. Communications in Statistics – Theory and Methods, 37(10):1671-1686.

Kazemzadeh, R.B., Noorossana, R., and Amiri, A. (2009) Monitoring polynomial profiles in quality control applications. International Journal of Advanced Manufacturing Technology, 42:703-712.

Kazemzadeh, R.B., Noorossana, R., and Amiri, A. (2010) Phase II monitoring of autocorrelated polynomial profiles in AR(1) processes. Transaction E: Industrial Engineering, 17(1): 12-24.

Kim, K., Mahmoud, M.A., and Woodall, W.H. (2003) On the monitoring of linear profiles. Journal of Quality Technology, 35:317-328.

Lai, T.-L. (1995). “Sequential Changepoint Detection in Quality Control and Dynamical Systems”, J. R. Statist. Soc. B, 57(4):613-658.

Leandro, G., Rolando, N., and Gallus, G. (2005). “Monitoring Surgical and Medical Outcomes: the Bernoulli Cumulative SUM Chart. A Novel Application to Assess Clinical Interventions”, Postgrad Med J, 81:647-652.

Liang, K.-Y., and Zeger, S. L., Longitudinal data analysis using generalized linear models, Biometrika, vol. 73, no. 1, pp. 13-22, 1986.

Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C., and Gallivan, S. (1997). “Monitoring the Results of Cardiac Surgery by Variable Life-adjusted Display”, The Lancet, 350(18):1128-30.

Mahmoud, M.A., and Woodall, W.H. (2004) Phase I analysis of linear profiles with calibration applications. Technometrics, 46(4):380-391.

Mahmoud, M.A., Parker, P.A., Woodall, W.H., and Hawkins, D.M. (2007) A change point method for linear profile data. Quality and Reliability Engineering International, 23:247-268.

Milo A Puhan, Madlaina Scharplatz, Thierry Troosters and Johann Steurer, Respiratory rehabilitation after acute exacerbation of COPD may reduce risk fro readmission and mortality – a systematic review, Respiratory Research 2005, 6:54.

Myers, R. H., Montgomery, D. C., and Vining, G. G., Generalized Linear Models−with Applications in Engineering and the Sciences, John Wiley&Sons, New York, 2002.

Neal, R. M. (2003) “Slice Sampling”, The Annals of Statistics, 31:705-741.

Noorossana, R., Saghaei, A., and Amiri, A. (2011) Statistical analysis of profile monitoring. John Wiley & Sons, Hoboken, NJ.

Noorossana, R., Vaghefi, A., and Dorri, M. (2011) Effect of non-normality on the monitoring of simple linear profiles. Quality and Reliability Engineering International, 27:425-436.

Novick, R. J., Fox, S. A., Stitt, L. W., Forbes, T. L., and Steiner, S. (2006). “Direct Comparison of Risk-adjusted and Non-risk-adjusted CUSUM Analyses of Coronary Artery Bypass Surgery Outcomes”, The Journal of Throracic and Cardiovascular Surgery, 132(2):386-391.

O. Grigg and V. Farewell, An overview of risk adjusted charts, J.R. Statist Soc A (2004)167, Part 3, pp 523-539.

Omar Hassan, David O Meltzer, Shimon A Shaykevich, Chaim M. Bell, Peter J. Kaboli, Andrew D. Auerbach, Tosha B. Wetterneck, Vineet M. Arora, James Zhang and Jeffrey L. Schnipper, Hospital Readmission in General medicine patients: A prediction model, J Gen Intern Med 25 (3): 211-219.

Pedro Almagro, Bienvenido Barreiro, Anna Ochoa de Echaguen, Salvador Quintana, Monica Rodriguez Carballeira, Jose L. Heredia and Javier Garau, Rsik factors for hospital readmission in patients with chronic obstructive pulmonary disease, respiration 2006; 73:311-317.

Phaladiganon, P., Kim, S.B., Chen, V., Baek, J.-G., and Park, S.-K. (2011). “Bootstrap-based T2 Multivariate Control Charts”, Communications in Statistics – Simulation and Computation, 40(5):645-662.

Pilcher, D. V., Hoffman, T., Thomas, C., Ernest, D., and Hart, G. K. (2010). “Risk-adjusted Continuous Outcome Monitoring with an EWMA Chart: Could It Have Detected Excess Mortality among Intensive Care Patients at Bundaberg Base Hospital?”, Critical Care and Resuscitation, 12(1):36:41.

Poloniecki J, Valencia O, Littlejohns P. Cumulative Risk Adjusted Mortality Chart for Detecting Changes in Death Rate: Observational Study of Heart Surgery. British Medical Journal 1998; 316: 1697-1700.

Qiu, P., and Hawkins, D. (2001) A Rank-based multivariate CUSUM procedure. Technometrics, 43(2):120-132.

Qiu, P., and Hawkins, D. (2003) A nonparametric multivariate cumulative sum procedure for detecting shifts in all directions. Journal of the Royal Statistical Society. Series D, 52(2):151-164.

Ravi P Kiran, Conor P Delaney, Anthony J Senagore, Malcolm Steel, Thomas Garafalo, Victor W Fazio, Outcomes and prediction of hospital readmission after intestinal surgery, The American college of surgeons 2004.

Robert D. Stewart, Christian T. Campos, Beth Jennings, Scott Lollis, Sidney Levitsky, Stephen J. Lahey, Predictors of 30 day hospital readmission after coronary artery bypass, The society of Thoracic surgeons 2000, 70:169-74.

Robert, C.P., and Casella, G (2004) Monte Carlo statistical methods, 2nd edition, Springer, New York, NY.

Ross, G.J., Tasoulis, D.K., and Adams, N.M. (2011) Nonparametric monitoring of data streams for changes in location and scale. Technometrics, 53(4):379-389.

Smith, D. M., Giobbie-Hurder, A., Weinberger, M., Oddone, E. Z., Henderson, W. G., Asch, D. A., Ashton, C. M., Feussner, J. R., Ginier, P., Huey, J. M., Hynes, D. M., Loo, L., and Mengel, C. E., Predicting non-elective hospital readmissions: a multi-site study, Journal of Clinical Epidemiology, vol. 53, pp. 1113-18, 2000.

Stefan H. Steiner, Richard J.Cook, Vern T Farewell, Tom Treasure, MOnitroing surgical performance using risk adjusted cumulative sum charts, Bisotatistics (2000) 1,4,pp.441-452.

Steiner, S. H., Cook, R. J., Farewell, V. T., and Treasure, T. (2000). “Monitoring Surgical Performance using Risk-adjusted Cumulative Sum Charts”, Biostatistics, 1(4):441-452.

Tze-Pin Ng, Mathew Niti, Wan Cheng Tan, Zhenying Cao, Kian-Chung Ong and Philip Eng, Depressive Symptoms and Chronic Obstructive Pulmonary disease, Arch Intern Med, 2007; 167:60-67.

Vaghefi, A., Tajbakhsh, S.D., and Noorossana, R. (2009) Phase II monitoring of nonlinear profiles. Communications in Statistics: Theory and Methods, 18:1834-1851.

Victor A. Ferraris, Suellen P. Ferraris, R. Christopher Harmon, Boyd D. Evans, Risk Factors for early hospital readmission after cardiac operations, The journal of thoracic and cardiovascular surgery, august 2001.

Walker, E., and Wright, S.P. (2002) Comparing curves using additive models, Journal of Quality Technology, 34:118-129.

Williams, J.D., Birch, J.B., Woodall, W.H., and Ferry, N.M. (2007) Statistical monitoring of heteroscedastic dose-response profiles from high-throughput screening. Journal of Agricultural, Biological and Environmental Statistics, 12:216-235.

Williams, J.D., Woodall, W.H., and Birch, J.B. (2007) Statistical monitoring of nonlinear product and process quality profiles. Quality and Reliability Engineering International, 23:925-941.

Woodall, W. H. (2006). “The Use of Control Charts in Health-Care and Public-Health Surveillance”, Journal of Quality Technology, 38(2):89-104.

Woodall, W.H. (2007) Current research on profile monitoring. Produção, 17(3), 420-425.

Woodall, W.H., Spitzner, D.J., Montgomery, D.C., and Gupta, S. (2004) Using control charts to monitor process and product quality profiles. Journal of Quality Technology, 36(3), 309-320.

Y.Chen, Q.Li and H.Johansen, Age and sex variations in hospital readmission for COPD associated with overall and cardiac comorbidity, The international journal of Tuberculosis and Lung Disease 13(3): 394-399.

Yea-Jyh Chen and Georgia L. Narsavage, Factors related to Chronic Obstructive Pulmonary Disease readmission in Taiwan, West L Nurs Res 2006 28:105.

Yehuda Kariv, Wei Wang, Anthony J. Senagore, Jeffrey P. Hammel, Victor W. Fazio, Conor P. Delaney, Multivariate analysis of factors associated with hospital readmission after intestinal surgey, The American Journal of Surgery 191(2006) 364-371.

Zeng, L., and Zhou, S. (2011). “A Bayesian Approach to Risk-adjusted Outcome Monitoring in Healthcare”, Statistics in Medicine, 30(29):3431-3446.

Zhenying Cao, Kian Chung Ong, Philip Eng, Wan Cheng Tan and Tze Pin Ng, Frequent hospital readmissions for acute exacerbation of COPD and their associated factors, Respirology (2006) 11, 188-195.

Zou, C., and Tsung F. (2011) A multivariate sign EWMA control chart. Technometrics, 53(1):84-97.

Biographical Information

Smriti Neogi graduated from Nirma University of Science and Technology, Gujarat, India with a Bachelor of Mechanical Engineering in 2001. She worked for a company called Larsen and Tuobro as Design Engineer until 2005. She received her Master degree in Industrial Engineering in 2007 from University of Texas at Arlington. She worked for a company called Cummins as Process Manufacturing Engineer for three years. She started her PhD in Industrial Manufacturing Systems Engineering Dept at the University of Texas at Arlington in 2010 under the supervision of Dr Li Zeng. Her dissertation topic is “Phase I Monitoring with Application in Manufacturing and Healthcare”.