STATISTICAL QUALITY METHODS TO MONITOR AND TRANSFORM HEALTHCARE DATA

A Dissertation in
Industrial Engineering

by
Min-Jung Kim

© 2012 Min-Jung Kim

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012
The dissertation of Min-Jung Kim was reviewed and approved * by the following:

Harriet Black Nembhard  
Professor of Industrial Engineering  
Dissertation Advisor  
Chair of Committee

Rhonda Belue  
Associate Professor of Health Policy and Administration

Paul Griffin  
Professor of Industrial Engineering  
Peter and Angela Dal Pezzo Head of Industrial and Manufacturing Engineering

Ling Rothrock  
Associate Professor of Industrial Engineering

Li Wang  
Assistant Professor of Public Health Sciences

*Signatures are on file in the Graduate School
ABSTRACT

The last decade has presented critical health challenges, ranging from the threat of new infections, to the emerging burdens of chronic disease. As experienced in the H5N1 avian influenza epidemic in 2007 and the H1N1 influenza pandemic in 2009, infectious diseases have emerged as a significant threat to global health. Furthermore, the rise in the prevalence of chronic conditions across all socioeconomic classes has made these conditions the leading causes of death and disability. These and other similar challenges have led to a growing interest in leveraging healthcare data better to facilitate disease prevention and control and better health care. In this dissertation, three statistical quality control (SQC) based methods are proposed to deal with some of the distinguishing characteristics in healthcare data.

The first method integrates SQC with forecasting methods to support syndromic surveillance in health monitoring. Health data are likely to be non-stationary, non-normal, or autocorrelated, which calls for caution in the application of traditional SQC approaches. Furthermore, in order to address the need to track movements of the disease progression/dispersion as well as obtain a timely and sensitive signal for disease change, the method proposes pretreatment steps using forecasting based methodologies (regression and time-series models) that strengthen in modeling baseline patterns of disease. The pretreatment steps are applied to health data in such a way that results in independently and normally distributed observations so that SQC procedures can be appropriately applied. The proposed health monitoring framework provides a customized procedure for dealing with various statistical characteristics of health data which lead to use satisfy general assumptions of SQC. As an example of influenza syndromic surveillance, we
monitor the weekly influenza-like-illness (ILI) incidence data and weekly over-the-counter and prescribed medication sales related to the ILI symptoms. Statistical analysis of two sources of data shows that they are non-stationary processes with a high level of autocorrelation which is a barrier to the direct use of traditional SQC. The application of our monitoring framework to the multiple streams of ILI and drug sales results in earlier warning than the official announcement of outbreak.

The second method combines two multivariate SQC approaches – the multivariate self-starting exponentially weighted likelihood ratio (MSS-EWLR) chart and the multivariate change-point (MCP) chart – for the complexity found in health monitoring applications where there is interest in two or more quality characteristics that may be correlated and where in-control process data is limited. As an example of its application in monitoring chronic disease, systolic blood pressure, diastolic blood pressure and mean arterial pressure are monitored for assessing hypertension and its related complications. Our approach allows the monitoring of both mean and variability of healthcare data in a single chart. The combination of the two SQC approaches provides better performance enabling the detection of subtle changes that can be masked by unsuspected change in the correlation among quality characteristics by monitoring both mean and variability. Also, the combined approach has the added benefit of working without *a priori* parameter information of a baseline.

The third method combines time-series analysis and regression to transform cross-sectional data into repeated measurement data describing individual patient profiles that can be monitored with SQC tools. The majority of accessible national surveys are based on cross-sectional data that depict a snapshot of health conditions in the population. Although national surveys are
performed periodically with same individuals, identifiers are not provided to protect participant privacy and confidentiality. This profile data estimation (PDE) method will allow the performance of proposed health monitoring methods to be demonstrated with real clinical data that is publicly available. This may ease the burden of validating new methodologies. Moreover, it may result in a better, more rapid understanding of the relationship among multiple data characteristics that explain disease progression and dispersion over a complex disease process within and between patients.

The methods developed here link quality engineering research in the field of industrial engineering with health monitoring systems research in the field of healthcare. Enhanced healthcare data monitoring will contribute to the goal of ensuring that care teams and relevant organizations receive accurate, timely, and up-to-date information about given health situations so that they can determine the appropriate intervention to undertake. In this sense, the work is a part of the broader partnership between healthcare and its associated clinical discoveries, and the engineering systems that help translate the discoveries into action.
# TABLE OF CONTENTS

ABSTRACT ........................................................................................................................... iii
LIST OF TABLES .................................................................................................................... ix
LIST OF FIGURES .................................................................................................................. x
ACKNOWLEDGEMENTS ....................................................................................................... xii

CHAPTER 1 INTRODUCTION ............................................................................................... 1
  1.1 Background .................................................................................................................... 1
    1.1.1 The Partnership between Healthcare and Engineering ........................................... 1
    1.1.2 Growing Importance of Health Monitoring Systems ............................................. 3
  1.2 Motivation ..................................................................................................................... 6
  1.3 Research Objectives ..................................................................................................... 10
  1.4 Outline of the Dissertation ......................................................................................... 12

CHAPTER 2 FUNDAMENTAL APPROACHES IN HEALTH MONITORING .................... 15
  2.1 Regression ..................................................................................................................... 15
    2.1.1 Trigonometric Methodology .................................................................................. 16
    2.1.2 Poisson Regression with Logarithm Scales ............................................................ 16
    2.1.3 Analogue Methodology ......................................................................................... 17
  2.2 Time-Series Models ..................................................................................................... 19
    2.2.1 ARIMA Time Series Models ............................................................................... 19
    2.2.2 SARIMA Time Series Models ............................................................................. 21
    2.2.3 VARIMA Time Series Models ............................................................................. 22
  2.3 Statistical Quality Control (SQC) ................................................................................ 23
    2.3.1 Traditional Univariate SQC .................................................................................. 25
    2.3.2 Traditional Multivariate SQC ................................................................................ 27
    2.3.3 SQC Applied to Autocorrelated Data .................................................................. 28
    2.3.4 Phase I and Phase II ............................................................................................. 29
  2.4 Change-Point Chart ..................................................................................................... 31
CHAPTER 3 A SYNDROMIC SURVEILLANCE SYSTEM FOR CLINICAL AND NON-CLINICAL HEALTH DATA ................................................................. 35
3.1 Introduction .............................................................................................. 35
3.2 Literature Review .................................................................................... 39
3.3 Method ..................................................................................................... 43
3.4 Data .......................................................................................................... 45
  3.4.1 Drug Sales Data .................................................................................. 46
  3.4.2 Influenza-Like-Illness (ILI) Data ........................................................ 45
3.5 Result ........................................................................................................ 47
  3.5.1 Univariate Change-Point Chart (ILI incidence at national level) .......... 49
  3.5.2 Multivariate Change-Point Chart (ILI incidence in 22 regions and 20 classes of drug sales) ................................................................. 50
  3.5.3 Discussion .......................................................................................... 52

CHAPTER 4 UNKNOWN PARAMETER MULTIVARIATE STATISTICAL QUALITY CONTROL MODELS TO MONITOR CHRONIC DISEASE ........................................... 58
4.1 Introduction ............................................................................................... 58
4.2 Literature Review .................................................................................... 61
  4.2.1 A General Framework for Simultaneous Monitoring ......................... 61
  4.2.2 A General Framework for Unknown Parameter Multivariate Control Charts .......................................................... 63
4.3 Method ..................................................................................................... 65
  4.3.1 Control Chart Construction ................................................................. 65
  4.3.2 Multivariate Self-Starting Exponentially Weighted Likelihood Ratio (MSS-EWLR) Chart ............................................................ 66
  4.3.3 Multivariate Change-Point (MCP) Chart ............................................ 68
4.4 Data: Blood Pressure Monitoring Case Study ......................................... 69
4.5 Results ..................................................................................................... 71
  4.5.1 Case Study ......................................................................................... 71
  4.5.2 Simulation Study: Comparisons of MSS-EWLR and MCP ................. 77
  4.5.3 Discussion .......................................................................................... 80
CHAPTER 5 ESTIMATING REPEATED MEASUREMENT DATA FOR HEALTH CONDITIONS ........................................................................................................................................... 85
5.1 Introduction ................................................................................................................................................................................... 85
5.2 Literature Review ...................................................................................................................................................................... 88
5.3 Data .......................................................................................................................................................................................... 92
5.4 Methods ................................................................................................................................................................................... 93
  5.4.1 Creating a Consolidated Dataset ................................................................................................................................. 95
  5.4.2 Sorting Data File by Variables of Research Interest ..................................................................................................... 97
  5.4.3 Regression Modeling (Age-Associated Physiological Trend) .......................................................................................... 97
  5.4.4 Creating a BP Profile for One Person ............................................................................................................................. 101
5.5 Results .................................................................................................................................................................................... 104
5.6 Discussion ................................................................................................................................................................................ 105

CHAPTER 6 CONTRIBUTIONS AND FUTURE RESEARCH ...................................................................................... 109
6.1 Contributions ......................................................................................................................................................................... 109
  6.1.1 Integrated Health Monitoring System with SQC, Regression Modeling, and Time Series Modeling .................................................. 109
  6.1.2 Improvement of Health Monitoring SQC Methods: MCP and MSS-EWLR ......................................................... 111
  6.1.3 Creating Synthetic Clinical Profiles from Cross-Sectional survey ........................................................................... 114
  6.1.4 Multiple Streams of Health Data: Clinical and Non-clinical Data ............................................................................... 115
6.2 Future Research ....................................................................................................................................................................... 116
  6.2.1 Data Collection Efforts for Health Monitoring: Non-clinical Data ............................................................................. 116
  6.2.2 Development of SQC Methods following Other Statistical Distributions ................................................................. 119
  6.2.3 Integration of SQC and Variability Research for Accurate Health Diagnosis .............................................................. 119
  6.2.4 Research on the Effect of Interventions in Chronic Disease Surveillance ................................................................. 121

REFERENCES ............................................................................................................................................................................ 122
VITA ........................................................................................................................................................................................ 149
LIST OF TABLES

Table 3.1. Twenty classes of drug sales likely to be purchased or prescribed in an ILI context.. 46
Table 4.1. First few cases of the BP Profile data showing the self-starting procedures and MSS-EWLR statistics................................................................. 72
Table 4.2. The $G_{\text{max}}$ statistics of the MCP chart and its corresponding change point (CP) of the BP profiles. (The rows in bold indicate where the $G_{\text{max}}$ exceeded control limit, h = 6.676.) 74
Table 5.1. Number of non-diabetic participants by age and BMI classification in NHANES (1999-2008)... ......................................................................................... 96
Table 5.2. Means and standard errors (SE) of SBP, DBP, and HbA1c by age and obesity categories in the non-diabetic population. (Source: NHANES 1999-2008) ................. 99
Table 5.3. SBP and DBP levels for normal-weight, over-weight and obese cohorts. ............ 104
LIST OF FIGURES

Figure 3.1. Syndromic surveillance helps prevent severe illness by treating patients with initial symptoms earlier than traditional detection. Here $t$ is the time between detection by syndromic surveillance and detection by traditional methods. 36

Figure 3.2. ILI incidence (source: French Sentinel Network, www.sentiweb.org) and several classes of drugs likely to be purchased or prescribed in ILI context (source: IMS France). 38

Figure 3.3. A change-point based surveillance scheme. 44

Figure 3.4. French Sentinel Network (www.sentiweb.com) Results. 47

Figure 3.5. Comparison between observed data, non-epidemic thresholded data and forecasted values obtained using the method of analogues. 48

Figure 3.6. ACF graph of residuals between $y_{\text{thresh}}$ and $y_{\text{an}}$ (before ARMA filtering). 49

Figure 3.7. Probability plots after ARMA filtering. 49

Figure 3.8. Univariate change-point applied to transformed ILI incidence (national) to detect epidemic onset. (ILI incidence data: solid line, change-point threshold: dashed line) 50

Figure 3.9. Multivariate change-point applied to transformed ILI incidence (22 regions) to detect epidemic onset. (ILI incidence data: solid line, change-point threshold: dashed line) 51

Figure 3.10. Multivariate change-point applied to transformed drug sales (20 classes of medications) to detect epidemic outbreak. 52

Figure 4.1. Ambulatory monitoring outcomes for SBP, DBP and MAP. 70

Figure 4.2. MCP (left) and MSS-EWLR (right) for ambulatory monitoring outcomes. 75

Figure 4.3. The effect of mean shift: ARL OOC performance study of the MSS-EWLR (Left) and MCP (Right) with different $\tau$ (20, 50, 100, 200) in the mean shift change: $\delta = (0.5, 1, 2, 4)$. Each line represents a specific mean shift, starting with 0.5 at the top and 4 at the bottom. 78

Figure 4.4. The effect of shift on covariance shift: ARL OOC performance study of the MSS-EWLR (Left) and MCP (Right) with different $\tau$ (20, 50, 100, 200) in covariance shift changes: $\rho = (0.2, 0.4, 0.6, 0.8)$. Each line represents a specific covariance shift, starting with 0.2 at the bottom and 0.8 at the top. 79
Figure 4.5. The effect of a shift in mean vector coupled with a shift in covariance matrix: ARL OOC performance study of the MSS-EWLR (Left) and MCP (Right) with different \( \tau \) (20, 50, 100, 200) in mean and covariance shifts: \((\delta, \rho) = [(0.5, 0.6), (0.5, 0.2), (1, 0.6), (1, 0.2)]\). Each line represents a specific coupled mean and covariance shift, starting with (0.5, 0.6) at the top and (1, 0.2) at the bottom.  

Figure 5.1. Overview of the profile data estimation (PDE) methodology as applied to blood pressure (BP).  

Figure 5.2. Mean value of physiological variables (SBP, DBP, HbA1c) in the non-diabetic population.  

Figure 5.3. BP profiles of 30 non-diabetic patients in who are normal-weight/over-weight/obese.
ACKNOWLEDGEMENTS

I would like to express my deepest appreciation and respect to my advisor, Dr. Harriet Black Nembhard for her constant guidance throughout the whole period of my doctoral studies. I sincerely believe that studying and working under her supervision have yielded the largest achievements in my career, and I am glad that I have had that opportunity.

I am also grateful to Dr. Paul Griffin, Dr. Ling Rothrock, Dr. Rhonda Belue, and Dr. Li Wang for being on my doctoral committee. I appreciate their generosity in sharing their knowledge, experience, and time while serving on my committee.

I appreciate Dr. Susan Griffin, Dr. Marija Jankovic and Dr. Eralp Dogu for enlightening and enthusiastic discussions with me on this research.

I wish to express my sincere thanks to all my friends and colleagues. I would like to thank to Dr. Cumpol Yuangyai, Dr. MonifaVaughn-Cooke, Hyojung Kang, Yonmi Jung, Melinda Green, and Aurore-Laetitia Mata for our discussions that took place in the Quality Engineering and System Transitions (QUEST) lab. I also owe thanks to many Korean IE students in State College who have supported me through conversations about research, life, and future.

I wish to extend my deep appreciation to my family. I would like to thank my parents for their consistent love, support, and sacrifice. I am thankful to my brother, sister, brother-in-laws, and sister-in-laws for encouraging me all the time. I would like to express my sincere gratitude to my father-in-law and mother-in-law for their encouragement and understanding. Lastly and most importantly, I would like to express my heartfelt appreciation and love to my husband, Insoo. He has generously and consistently provided me the incredible support, encouragement, patience
and love. The best experience of this graduate school period was the birth of our daughter, Sophia, who dramatically changed and enriched our life.
CHAPTER 1
INTRODUCTION

This dissertation links quality engineering research in the field of industrial engineering with health monitoring systems research in the field of healthcare. This chapter provides the background, motivation, and objectives of this research investigation.

1.1 Background

1.1.1 The Partnership between Healthcare and Engineering

The United States leads the world in medical science and technology, defining the cutting edge in most fields of clinical research and practice worldwide (AdvaMed, 2004; Kern, Thomas, & Hughes, 2009; Light, 2008; Mechanic, 2009; NAE, 2003; Pidgeon, Harthorn, Bryant, & Rogers-Hayden, 2008). The U.S. market for healthcare services has supported this focus by rewarding innovation in medical procedures and interventions, and drugs, devices, and equipment linked to them with relatively little regard for cost (Bodenheimer, Chen, & Bennett, 2009; Fisher, Bynum, & Skinner, 2009; Marmor, Oberlander, & White, 2009; NAE. & IOM., 2005; Orszag & Emanuel, 2010; Sutherland, Fisher, & Skinner, 2009). As a result, the U.S. healthcare system can provide high quality, highly specialized care for some individuals. However, less than half of patients in the U.S. benefit from the established “best practice”, and the cost of healthcare per person in the country is the highest among the top 21 industrialized nations (IOM, 2001, 2004; Naumann, Dellinger, Zaloshnja, Lawrence, & Miller, 2010; Zhang, et al., 2009).
The gap between the rapidly advancing medical knowledge base and its applications to patient care comes to the forefront as crises in safety, quality, cost and access seriously threaten the health of the American population (Driebe & McDaniel, 2005; Reid, 2005). A National Academy of Engineering and Institute of Medicine report (2010) suggested that the main cause of these crises can be explained by the complexity of the American healthcare system. Yet, little effort has been expended to effectively engage the healthcare system as a whole and improve its operations. The complexity of the healthcare system stems from six interconnected levels – the patient, the population, the team, the organization, the network, and the political and economic environment (PPTONE) – that are comprised of several components such as multiple clinicians within a clinic, or multiple clinics in a hospital (Rardin, 2007). In addition to the complex structure of the healthcare system, the dynamic associations between and within levels and their corresponding components yield high degrees of complexity.

To address system complexities and respond to challenges and hitherto missed opportunities, a partnership between engineering and healthcare is being formed that capitalizes on the potential of system engineering tools to improve the healthcare system. Controlling a complex system requires a clear understanding of performance expectations and the operating parameters for meeting those expectations. System engineering tools have been used to improve the quality, efficiency, and operations of many complex systems. In healthcare, these tools and approaches can support research and development across PPTONE levels. In particular, the direct interaction between the engineering community and healthcare community has been advised to facilitate the development of productivity measures and monitoring systems (NAE, 2003; NIH, 2007).
1.1.2 Growing Importance of Health Monitoring Systems

There is growing interest in health monitoring systems based on the threat of new infections, the increased potential for bioterrorism and emerging burdens of chronic disease. In recent years, the global community has had to respond to several major infectious disease emergencies – e.g., anthrax attacks in 2001, H5N1 avian influenza epidemics in 2007 and H1N1 influenza pandemics in 2009. As a result, public health authorities acknowledged the need for improved surveillance in order to detect emerging infectious disease and syndromes earlier than would the standard reporting of laboratory confirmed cases, in order to mitigate impact of the emerging outbreaks by implementing intervention strategies in a timely way.

For instance, the H1N1 epidemic could have caused a major pandemic with millions of infections and deaths if it had not been identified and controlled by public health practitioners. Still the H1N1 epidemic was only recognized several months after its emergence, and many infected patients and deaths occurred prior to its acknowledgment (Chan, 2009). This implies a need to design new syndromic surveillance systems, which refers to a type of public health surveillance with the specific goal of detecting disease outbreaks earlier than conventional reporting of laboratory confirmed cases. Historically, public syndromic surveillance has been geared towards monitoring natural morbidity and mortality associated with infectious diseases. There is an important connection between syndromic surveillance and health monitoring systems in order to detect abnormal signals that indicate disease dispersion in accurate and timely ways.

Another concern is that the prevalence of chronic conditions in all socioeconomic classes across the world has made these conditions the leading causes of death and disability. They now account for almost 60% of all deaths and 43% of the global burden of disease (WHO, 2009).
What is worse is that the increasing tendency of chronic disease will be projected to rise to 74% of all deaths and 60% for the global burden for the chronic disease by 2020. In the U.S., chronic disease affects the quality of life of 90 million U.S. residents and the cost of medical care for persons with chronic disease accounts for 70% of total medical care expenditures (Paez & Hwang, 2009; CDC, 2012).

These figures suggest that there will be an increase in chronic disease with the consequent burdens on society, the healthcare systems as well as individual patients. The appropriate response is that chronic disease management should focus on integrated prevention and controlling the risk factors of the disease. It requires both effective treatment and ongoing monitoring based on the analysis and interpretation of the periodic measurements that can guide the management of chronic or recurrent conditions. Monitoring health outcomes can help patients reach goals by allowing their health status to be diagnosed more accurately, by detecting the potential risk of disease progression in a timely way, and by enabling patients to be active participants in their own care.

The adoption of electronic health records (EHR) presents another dimension of challenge. The EHR system was originally designed to collect and analyze patient health data in real time in order to facilitate patient-centered treatment. Data collection occurs during patient visits and the medical data are stored along with personal information. An ongoing collection of medical records for an individual patient can be used to display health outcomes in a useful format. A visual summary of patient health records helps clinicians understand and interpret the data in order to make accurate diagnoses by comparing previous medical records to basic target values. In order to turn patient data into information that can guide treatment, it is necessary to discern
disease patterns and detect abnormalities that signal disease changes. Data monitoring is what transforms collected medical data into useful information by incorporating treatment guidelines in an intelligent and timely way.

The EHR system also can aggregate information in order to form patient subpopulations, thereby generating medical information and knowledge in a quantifiable way. Data pertaining to patient medical record histories and medical information can be organized by disease, demographics, family status, functional status and other research objectives. At no implementation cost in terms of money and time, the EHR system can provide patient medical information by recognizing patterns in care, outcomes, and practices. These rich pools of data can be used to evaluate public health concerns or for population health surveillance. Careful statistical monitoring analysis and observation of patients and their care programs is essential to adequately understanding and managing the application of medical knowledge generated from individuals to groups. While we do not specifically explore EHR design and integration, we recognize the impact it has had in creating new demand for the development of health monitoring systems that incorporate sophisticated statistical methods and engineering computational tools.

These challenges have led to a growing interest in improving health monitoring systems by leveraging healthcare data better to model and describe the input, process, output response that impact disease control and prevention. This dissertation addresses the development of statistical quality control (SQC) methods that can be used to measure, evaluate, monitor and control the output using healthcare data.
1.2 Motivation

Historically, health monitoring has been focused on monitoring the naturally occurring morbidity and mortality associated with infectious disease and health outcomes corresponding to chronic diseases. Almost all health monitoring approaches are retrospective, in the sense that they describe baseline health related data from the recent past and predict particular expected outcomes. A retrospective approach can help to evaluate the performance of a health monitoring system, but it cannot sufficiently achieve the aim of detecting important changes in the disease process as soon as possible because of the limited predictive power of changes in disease.

To deal with this issue, SQC can be used to provide the benefit of timely continuous feedback by monitoring the stream of health data repeatedly and periodically in a prospective way. SQC methods are statistical tools for monitoring quality and process parameters and alerting when there is an indication that those parameters have changed. After something unusual is noted in the health data pattern, health practitioners and professionals will attempt to confirm whether a disease is progressing or dispersing, similar to a search for assignable causes pursuant to an SQC control chart signal. Thus, SQC have been used as a prospective control tool in health monitoring in order to accurate, timely and up-to-date information about the given health situations so that they can determine the appropriate intervention to undertake.

SQC was originally developed for industry applications; texts such as Montgomery (2009) trace the history from Walter Shewhart’s seminal work in the 1920s at Bell Labs to automotive manufacturing and chemical processing. The main objective of SQC is to identify unusual patterns of variation and evidence of special causes when process values fall outside the control limits. The notion of aberration detection in health monitoring is similar to the idea in SQC.
where an out-of-control signal warrants further investigation of special causes (Bueher, et al., 2004). However there are still many difficulties with validating and accomplishing sensitive performance with SQC in health monitoring. Among these, we can summarize three points that must be overcome as follows.

1) **The characterization of the health data in terms of stationarity, autocorrelation, and distribution.** Although SQC methods have been used for healthcare data, there are some hurdles to its direct application. Though often not stated explicitly, the basic assumption of SQC methods is that observations are independent, identically distributed and typically normally distributed (or with a known parameter distribution). When SQC methods are applied in industrial settings where there is a more controlled environment, a target/nominal value, and close approximation to the normal distribution. Observations in health monitoring have various statistical characteristics in terms of stationarity, autocorrelation and identical distribution.

   For example, syndromic surveillance data are more likely to be non-stationary (i.e., reflect a seasonal pattern) and autocorrelated. Two methods that have evolved to meet this challenge – the early aberration reporting system (EARS) and the electronic surveillance system for the early notification of community-based epidemics (ESSENCE) – implemented exponential weighted moving average (EWMA) or cumulative sum (CUSUM) SQC methods in order to detect early aberrations for influenza-like illness and other disease (Hutwagner et al., 2005; Zhu et al., 2005). These methods typically estimate change points in a sequence of infectious disease events or rates (Benneyan, 1998; Carey, 2003; Woodall, 2006).
In addition, healthcare data related to chronic disease monitoring tend to exhibit high autocorrelation because of the dynamic properties on individual-level health outcomes. Further, given the goal of quick detection, methods are usually run on individual observations for which the normality assumption generally does not apply. Nevertheless, SQC methods have been used for monitoring variables at patient level such as blood pressure or blood sugar (Ryan, 2000; Wheeler, 2003).

The general approach that underlies the ability to overcome SQC’s basic assumption in terms of stationarity, independence and normality is statistical forecasting methods, namely with regression and time-series modeling. These methods are employed to precondition and transform the observed data in order to make it more conducive to SQC applications (Montgomery, 2009). The basic idea is to model the expected number of events by forecasting methods including terms of seasonal variations, dynamic properties of auto-correlation and cross-correlation in the model and then remove the explainable/known effects from the observations. Once observations are preconditioned by statistical forecasting methods, the residuals between the expected number and observed number can facilitate to be monitored using traditional SQC methods.

2) **Multiple quality characteristics that may be correlated and without known parameter information.** Much of the previous work to use SQC in health care has tended toward simpler applications by favoring univariate over multivariate monitoring, and by monitoring only location shifts rather than simultaneously monitoring location and scatter. However, many real health monitoring settings require tracking two or more quality characteristics that are usually correlated. When monitoring with traditional multivariate SQC procedures, a change in process
mean can be masked by unsuspected change in covariance structure or by a sudden change in the correlation among quality characteristics.

Furthermore, it may be difficult to gather sufficient in-control process data in order to obtain the parameter information required to define control limits. Most of the control practices assume that the parameter such as mean and variability are known \textit{a priori}. Here, the “plug-in” method is often used where a large data sequence is gathered while the process is believed to be in control. This technique can be acceptable in controlled industrial environments with clearly defined specifications or targets. However, when monitoring an individual patient, it is often unrealistic to gather sufficient historical data for the in-control process. Even if there are enough data to estimate the parameters, it cannot be guaranteed that the obtained parameters describe the in-control process.

3) \textbf{Limited sources of clinical profile data that can be used for monitoring.} Clinical data is needed to explain disease process within/between patients. Many epidemiologic studies have been based on repeated measurements data involving individual health status; however, such studies require large financial and time resources. It can be risky to invest money and time gathering repeated measurement data without some assurance of the outcome. Thus researchers are recommended to outline possible scenarios of their hypothesis and interpretations and try a pilot study with alternative dataset for validating the new trials, before expending resources to collect repeated measurement data.

Most of the alternative data that researchers can access are based on cross-sectional data from national surveys. These data depict a snapshot of the population health conditions and that are
designed to make county-level estimates of health status and health service utilization. Even if national surveys can be performed periodically with same persons, they do not provide individual identifiers to protect the privacy of the respondents’ information. This situation limits the analysis of the dynamic changes of individuals over time with available national survey data.

1.3 Research Objectives

The main purpose of health monitoring is to detect adverse symptoms in the body as well as new epidemics in public and to signal early warning for the changes in underlying disease patterns which can help understand overall situational awareness of disease activity. To address the timeliness and trustworthiness of information related to changes in disease patterns, SQC methods have been incorporated into the health monitoring system.

However, hurdles remain in validating and accomplishing sensitive performance with SQC methods in health monitoring. Based on the issues in using SQC methods to monitor and transform health data in the preceding section, the following research objectives have been formulated.

1) **To integrate SQC methods and statistical forecasting methods to overcome the basic assumptions of SQC in syndromic surveillance (Chapter 3).** Forecasting methods create restrictions because they are based on predicting the future value with a long sequence of past data under the assumption that the data pattern is consistent. This implies that there is no guarantee that the selected model will remain unchanged in the future, even if the forecast model fits well with the historical data. Therefore, we propose a circumspect integration of
SQC with forecasting methods to cope with the limitations and the assumptions inherent in each.

2) **To develop advanced multivariate SQC tools which enable to monitor multiple quality characteristics which are correlated each other in terms of mean and variability without *a priori* parameters information (Chapter 4).** In light of the challenges of SQC uses in actual health monitoring applications, we propose the integration of the multivariate self-starting exponentially weighted likelihood ratio (MSS-EWLR) chart and multivariate change point (MCP) chart to enable monitoring of both mean and variability in a single chart without *a priori* parameter information of the baseline. This advanced SQC approach provides better performance detecting subtle changes which can be masked by unsuspected change in the correlation among quality characteristics by monitoring both mean and variability.

3) **To create synthetic clinical health data profiles from cross-sectional survey data (Chapter 5).** In response to ongoing need of repeated measurement data of health outcomes, we suggest the profile data estimation (PDE) methodology to estimate the profiles of health outcomes. The PDE can relieve the burdens of validating new research by reducing time and data collection costs. Moreover, it may result in a better comprehension of patient health status, which ultimately supports a higher quality of care.
1.4 Outline of the Dissertation

In this dissertation, we will demonstrate how the engineering perspective can be applied in health monitoring systems. Chapter 2 presents a review of the literature on health monitoring methods. Many methods and algorithms for disease detection have been evaluated in order to obtain timely information about disease changes. The most common existing monitoring methods for disease dispersion or progression can be largely categorized into statistics-based forecasting methods and engineering control tools. The basic forecasting methods are regression and time series models; these can model the baseline patterns of disease. As an engineering control tool, SQC charts assume the role of detecting abnormalities in disease patterns with the aim of preventing and controlling disease.

Chapter 3 presents the integrated syndromic surveillance framework of multivariate SQC tools and time series models in order to monitor the combination of clinical data and non-clinical data. The notion is that non-clinical data may be available that provides an earlier evidence of sickness than clinical data. Furthermore, considering multiple streams of data allows the possibility of detecting abnormalities in inter-relationships among the variables as well as with each variable. To investigate these ideas, this study dealt with clinical data on 22 regions of influenza-like illness (ILI) weekly incidence and non-clinical data on over-the-counter (OTC) and prescribed weekly medication sales data. The outcome shows not only that the proposed multivariate monitoring method provides an improved ability to recognize warning signals, but also that non-clinical data (drug sales data) provide an indirect but earlier indication of influenza outbreaks. The proposed method has the advantage of being both practical and relatively simple.
to implement and could be easily adapted to develop adequate surveillance tools for specific diseases.

Chapter 4 provides the MSS-EWLR chart and MCP chart for tracking two or more quality characteristics where there may be a change in both the process mean and variability during the monitoring period. We applied these suggested two advanced SQC methodologies (MSS-EWLR and MCP) to the blood pressure (BP) monitoring in order to assess risk for hypertension as well as its related complications (such as cardiovascular disease, diabetes and heart related disease). Systolic BP, diastolic BP, and mean arterial pressure are monitored with these suggested two charts to detect unexpected changes in the physiological variable for assignable cause variation, such as disease onset or disease progression. Furthermore, these charts can overcome the need to define parameters (mean and covariance matrix) during in-control (IC) process because it incorporates a generalized likelihood ratio (GLR) test. Thus, the proposed method enables mean and variability to be monitored in a single chart, improving monitoring sensitivity for various shifts without long sequences.

Chapter 5 provides a methodology for creating synthetic patient and cohort blood pressure profiles. Most accessible national surveys are based on cross-sectional data that depict a snapshot of health conditions in the population and are designed to provide county-level estimates of health status and health service utilization. Even though national surveys are performed periodically with same individuals, the ability to analyze dynamic changes in individuals over time is limited because personal identifiers are omitted. To support the need for time series health outcome data, we propose the profile data estimation (PDE) methodology to estimate patient profiles based on health outcomes of chronic disease.
In Chapter 6, the contributions of this research are discussed. A general assessment of health monitoring systems is made by comparing conventional approaches to the approaches we have developed. Finally, directions for future research are suggested, including potential extensions to the dissertation work.
CHAPTER 2

FUNDAMENTAL APPROACHES IN HEALTH MONITORING

In order for health monitoring to be effective, statistics-based analysis and control charts are required for tracking the disease trends as well as process monitoring. These statistics based analysis and engineering tools improve the ability to detect abnormal disease patterns in health monitoring because they provide accurate and timely information of disease activity. This enables more effective treatment and may alleviate the impact of disease threat. The most common disease monitoring methods include regression, time-series, and statistical quality control (SQC). Each is discussed in the subsection of this chapter.

2.1 Regression

The goal of regression modeling is to obtain a mathematical model that can describe the relationship between an outcome variable and certain factors. Regression methods have been widely used in health monitoring, both for detecting outbreaks in surveillance systems on the basis of laboratory reports and for detecting abnormality in health-outcomes which lead to the disease progression. This approach can evaluate whether observed counts are greater than expected with respect to natural disease occurrence history.

The regression based approach differs from other areas of biostatistics, in that they are used primarily to obtain standardized residuals. The residuals can be evaluated from the reported
incidence data and baseline value for the occurrence of certain disease. The distribution of these residuals in the absence of an outbreak is then used to determine a threshold value.

2.1.1 Trigonometric Methodology

Collins (1932) defined an epidemic as “two consecutive weeks of incidence above epidemic threshold in the presence of documented influenza activity”. Following this definition of epidemic, Serfling (1963) suggested a trigonometric regression model for estimating excessive pneumonia and influenza mortality. The regression model fits the non-epidemic data and provides a predictive value for a non-epidemic level curve. During a pandemic period, the proportion of the population affected may suddenly increase. An appropriate function can be found by combining a linear trend describing a stable trend with the sine and cosine terms describing periodic changes. This forms an equation of the type:

\[ y_t = u + vt + \sum a_t \cos \theta + \sum b_t \sin \theta + \varepsilon_t. \]

Based on the trigonometric methodology, Pelat, Turbelin, Boelle, Lambert, and Valleron (2009) suggested an improved prospective analysis of weekly counts of gastrointestinal disease (2002-2007) by taking a three degree polynomial regression instead of the linear term of the original Serfling’s original trigonometric model in order to offer more flexibility.

2.1.2 Poisson Regression with Logarithm Scales

In health monitoring, data observed is numeric but in the form of counts. It is common to consider the number of new cases or incidences of specific disease occurring in a population
over a certain period of time. The aim of regression analysis in such instances is to model the dependent variable \((Y)\) as the estimate of outcome using some or all of explanatory variables \((X_1, X_2, ..., X_n)\). The Poisson regression can analyze and model these count data.

The Poisson regression has different constraints. Poisson regression assumes the response variable \(Y\) has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters as seen in

\[
\log_e (Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n.
\]

In other words, the typical Poisson regression model expresses the log outcome rate as a linear function of a set of predictors. Poisson regression has advantages for estimating the incidence of infectious disease and its related symptoms (e.g., ILI symptoms) in order to predict the baseline activities during non-epidemic seasons (Hu et al., 2007; Kuhn et al., 1994; Tobias et al., 2001; Zeger, 1988). Poisson regression has also been used in comparative epidemiologic studies of incidence and mortality in order to identify stronger associations with the selected risk factors, age and potential confounders (Dekker, et al., 2000; Del Rincón, Williams, Stern, Freeman, & Escalante, 2001; Ergin, Muntner, Sherwin, & He, 2004; Rosamond, et al., 1998). The statistical analysis of this risk is of practical use for epidemiologists and health authorities, who want to inform estimated values of the incidence and its risk or to decide whether a more detailed intervention is needed.

### 2.1.3 Analogue Methodology

The analogue method is a non-parametric approach that originated in meteorology (Lorenz, 1969). It is a popular tool for forecasting weather (Mullan & Thompson, 2006; Riordan &
Hansen, 2002; Timbal & McAvaney, 2001) and has found application in hydrology (Diomede, Nerozzi, Paccagnella, & Todini, 2006). Analogue methods involve predicting changes that may take place tomorrow by using a day in the past when conditions appear to have been similar i.e., an analog. The method suggests that a forecaster can predict that a situation in the current forecast will behave the same as it did in the past. The analogue method can be difficult to use because it is virtually impossible to find a perfect analog from a previous scenario. Even small differences between the current situation and previous analog can lead to significantly different results. However, as time passes and more data are collected, the chances of finding a closely aligned analog for the current weather situation should improve, and so should forecasts based on this method.

Viboud, Boëlle, Carrat, Valleron, and Flahault (2003) used the analogue method for epidemiologic data modeling, specifically to forecast influenza-like-illness (ILI) activity in France. This method is based on the idea that forecasts can be constructed thus: by taking all data corresponding to influenza illnesses and expressing them as the weighted sum of vectors selected from historical influenza time series that match current activities. This method has been shown to provide useful predictions based on a correlation between forecasted and observed values. The basic principle of the analogues method is to first define the vector of ILI incidence (i.e., the vector of the number of new ILI cases per unit of time during several consecutive weeks) at time, \( T \), to measure the pattern over the previous weeks, \( l \):

\[
X(T) = [I(T), I(T - 1), ..., I(T - l)],
\]

where, \( I(t) \) is the observed incidence at time, \( t \), and \( X(T) \) is the vector comprising \((l + 1)\) consecutive observations until \( I(T) \). A subsequent exhaustive search identifies the nearest \( v \)
“neighbors” in the past that closely match the incidence vector by computing and minimizing the distance criterion between the current observed vector of incidence, $X(T)$, and a given vector from the past, $X(t)$:

$$\text{dist} \left( X(T), X(t) \right) = \sum_{j=1}^{l} (I(T-j) - I(t-j))^2.$$ 

Finally, by forming the weight, $w$, from the weighted inverse distance of the closest neighbors of $v$, the future value of time period, $h$, is forecast using this relation:

$$F(T+h) = \sum_{1 \leq i \leq v} w^i \cdot I(T^i + h),$$

where, $T^i$ is the time of the nearest neighbor, $i$. The choice of constants is the result of performing the forecast by testing different values of $v$ (2 to 16 neighbors) and $l$ (0 to 10 weeks) and choosing the set of parameters that minimizes a cross-validation criterion based on the root mean square error (i.e., the values that provide the most accurate forecasts in the retrospective series).

### 2.2 Time-Series Models

Time series models are used to represent the correlation structure of the data. The correlation structure can be defined with auto-correlation or seasonality components. Both types of components are widely used in health monitoring applications.

#### 2.2.1 ARIMA Time Series Models

Auto-regressive integrated moving average (ARIMA) models (Box & Jenkins, 1970) have been used for detecting outbreaks of infectious disease (Choi, Atkinson, Karlson, & Curhan, 2005;
ARIMA models essentially consist of three components: an autoregressive (AR) series, a moving average (MA) series, and an integrated component (I).

AR regresses on itself, and a $p^{th}$ order autoregressive process AR($p$) satisfies the equation as follows:

$$X_t = \Phi_1 X_{t-1} + \Phi_2 X_{t-2} + \cdots + \Phi_p X_{t-p} + \varepsilon_t,$$

where, $\Phi_i$ are real constants and $\varepsilon_t$ represents white noise with mean $0$ and standard deviation $\sigma_\varepsilon$. Another representation of the AR($p$) can be given as

$$\Phi(B) X_t = \varepsilon_t,$$

where $\Phi(B) = 1 - \Phi_1 B + \Phi_2 B^2 + \cdots + \Phi_p B^p$ is the polynomial in the backshift operator with roots outside the unit circle.

An MA series of order $q$, MA($q$) is defined as

$$X_t = \varepsilon_t - \Theta_1 \varepsilon_{t-1} - \Theta_2 \varepsilon_{t-2} - \cdots - \Theta_q \varepsilon_{t-q},$$

$$X_t = \Theta(B) \varepsilon_t,$$

where $\Theta(B) = 1 - \Theta_1 B + \Theta_2 B^2 + \cdots + \Theta_p B^p$ with constant coefficient $\Theta_i$.

To achieve greater flexibility in fitting actual time series, both AR and MA terms can be included in the model. This leads to the mixed auto-regressive moving average (ARMA) model. A mixed ARMA process containing $p$ AR terms and $q$ MA terms is said to be an ARMA process of order ($p,q$). It is given by,

$$\Phi(B) X_t = \Theta(B) \varepsilon_t.$$ 

The importance of the ARMA process lies in the fact that it may often describe a stationary time series. In practice, the assumption of stationarity can be violated. To achieve a stationary model,
it is necessary to remove non-stationary sources of variation. If the observation is non-stationary, then it is advisable to check whether the difference of the series can be stationary. The first difference, that is, \( w_t = X_t - X_{t-1} = (1 - B)X_t \), or \( d \)-order differences, \( w_t = (1 - B)^d X_t \), may produce a stationary time series. We will further call \( X_t \) an ARIMA process of \( p,d,q \). It is \( d \)th difference, denoted by \( w_t = (1 - B)^d X_t \), produces a stationary ARMA\((p,q)\) process.

Hence an ARIMA\((p,d,q)\) can be written as

\[
\Phi(B)(1 - B)^d X_t = \Theta(B)\epsilon_t.
\]

### 2.2.2 SARIMA Time Series Models

Time series data may sometimes exhibit strong period patterns, which are often referred to as the time series having seasonal behavior, besides trend components. This mostly occurs when the data is taken in specific intervals, monthly, weekly, and so on. One way to represent such data through an additive model where the process is assumed to be composed of two parts is as follows:

\[
X_t = S_t + N_t,
\]

where \( S_t \) is the deterministic component with periodicity \( S \), and \( N_t \) is the stochastic component that may be modeled as an ARIMA process. When seasonal term \( (S_t) \) is observed with periodicity \( s \), we have \( S_t = S_{t+s} \) or \( S_t - S_{t-s} = (1 - B^s) S_t \). \( X_t \) then can be seen as a process with a predictable interval with some noise superimposed. Hence a general seasonal ARIMA model of orders \( p,d,q \times (P,D,Q) \) with period \( s \) is

\[
\Phi^*(B^s)\Phi(B)(1 - B)^d (1 - B^s)^D X_t = \delta + \Theta^*(B^s)\Theta(B)\epsilon_t.
\]
where \( p \) is the AR order, \( q \) the MA order, \( d \) is the number of differencing operations, and \( P, D, \) and \( Q \) are the corresponding seasonal orders.

The SARIMA model is an effective tool for capturing complex seasonal patterns as well as trend component. This forecast based method has been used to fit the population behavior for the infectious disease with seasonal effects. Reis and Mandl (2003) observe number of patients reporting to the ED with respiratory symptom for syndromic surveillance and define the baseline without the outbreak as SARIMA model. Nobre, Monteiro, Telles, and Williamson (2001) evaluate the SARIMA model for estimating case occurrence of two modifiable diseases (reported number of cases of hepatitis A and malaria for the United States).

### 2.2.3 VARIMA Time Series Models

In practical applications of forecasting problems, there is need to track more than one variable. One approach to deal with multiple variables is to define separate model with individual variables that can be used to forecast each variable. However, this approach can ignore the correlation among variables which diminishes forecasting ability. Alternatively, the multivariate time series modeling techniques can be used to remove the correlation structure among variables as well as auto-correlation structure in observations. Box and Jenkins (1970) and Hamilton (1994) introduced the multivariate time series model in vector ARIMA (VARIMA) models which can be successfully used in forecasting multivariate time-series.

Suppose that the vector time series \( X_t = (x_{1t}, x_{2t}, ..., x_{mt}) \) consists of \( m \) univariate time series. As in the univariate case, the stationary vector time series can be represented with a vector ARMA (VARMA) model given by
\[ \Phi(B)X_t = \Theta(B)\varepsilon_t, \]

where \( \Phi(B) = I - \Phi_1 B + \Phi_2 B^2 + \cdots + \Phi_p B^p \), \( \Theta(B) = I - \Theta_1 B + \Theta_2 B^2 + \cdots + \Theta_p B^p \), and \( \varepsilon_t \) represents the sequence of independent random vectors with \( \text{E}(\varepsilon_t) = 0 \) and \( \text{Cov}(\varepsilon_t) = \Sigma \).

Similarly, if the observations are non-stationary, the vector ARIMA (VARIMA) model can be represented as

\[ \Phi(B)D(B)X_t = \Theta(B)\varepsilon_t, \]

where \( D(B) = \text{diag}((1 - B)^{d_1}, (1 - B)^{d_2}, \ldots, (1 - B)^{d_m}) \).

2.3 Statistical Quality Control (SQC)

The basic theory of SQC for monitoring process stability is the control chart, first introduced by Dr. Walter A. Shewhart at Bell Laboratories in the 1920s. A SQC chart turns time-ordered data for particular quality characteristics into a picture that is easy to understand and react to when necessary. The basic model for a Shewhart chart of individual measurements assumes that the process operates with a constant mean \( \mu \) and the measurement at time \( t \) (\( Y_t \)) can be represented as

\[ Y_t = \mu + \epsilon_t \quad \text{for } t = 1,2,\ldots \]

where, the errors \( \epsilon_t \) is a random error from the process mean \( \mu \). The errors are typically assumed to be statistically independent in the derivation of the conventional control limits that are displayed at three standard deviations above and below an estimate for mean \( \mu \).

SQC is a branch of statistics that combines rigorous time series analysis methods with graphical presentation of data, often yielding insights into the data more quickly and in a way that is more understandable to lay decision makers. A typical SQC chart is displayed of four
factors: the centerline, upper-control limit, lower-control limit and the main statistic of the observation. The centerline represents the average valued of the quality characteristic corresponding to the in-control state. Two other horizontal lines, called the upper control limit and lower control limit, are chosen so that if the process is in control, nearly all of the sample points will fall between them. As long as the points of main statistic plot within the control limits, the process is assumed to be in control and no action is necessary, however, a point that plots outside of the control limits is interpreted as evidence that there are unusual shifts in a process and investigate and corrective actions are required to find and eliminate the causes for the out-of-control signal.

The primary objective of using control charts is to distinguish between common cause (chance) and special (assignable) causes of variations. The former is generic to any (stable) process, and its reduction requires action on the constraints of the process; whereas special-cause variation requires investigation to find the cause, and, where appropriate, action to eliminate it. The control chart is one of an array of quality-improvement techniques that can be used to deliver continual improvement. Before control charts are constructed, it is essential to have a clear aim and a clear plan of action regarding how special-cause data points will be investigated. It is also important to ensure that individuals involved in the project are made aware of the SQC approach to understanding variation. An out-of-control signal is triggered if there is sufficient evidence that the process has deviated from its normal operation because of an assignable cause; otherwise, the process is said to be in (statistical) control. Using the usual $3\sigma$ upper and lower limit, the probability of false alarm (type I error) is 0.0027, and the average-run-length is 370.
The notion of aberration detection in health monitoring is similar to the notion of detecting the observation corresponding to the special cause. The disease progression or dispersion is then investigated in order to determine if the aberration is in fact a sign of real disease exposure. SQC provides researchers and practitioners with a method for better understanding and communicating data from healthcare improvement efforts.

Over the past few decades, a number of SQC methods have been developed to detect changes of disease in health monitoring (Montgomery, 2009). These methods typically estimate change points in a sequence of disease events or time series of population rates. They can also determine the control limits for the behavior of a system. Several traditional SQC methods, such as the Shewhart chart, cumulative sum (CUSUM) chart, and exponentially weighted moving average (EWMA) are currently used to assist in assessing change or in-control behavior (Woodall, 2006) in health monitoring systems. And more advanced monitoring SQC methodologies have been used for more sensitive detection and for more practical uses in health monitoring; multivariate SQC methods when monitoring multiple variables, time-series based SQC methods when monitoring autocorrelated variables.

2.3.1 Traditional Univariate SQC

The Shewhart chart performs best in detecting a large step shift. For more sensitive detection ability of small shift which is the main performance of charting procedure, EWMA (Hunter, 1986; Roberts, 1959) and CUSUM (D.M. Hawkins & Olwell, 1998; Page, 1954) are developed. The cumulative sum (CUSUM) chart is one of the most common tools from statistical quality control adapted to monitor adverse events in health monitoring (Morton, et al., 2001; Woodall,
The original CUSUM was first proposed by Page (1954) as a visual technique for plotting the cumulative sum of deviations from the mean, placing emphasis on keeping the process on am rather than allowing it to drift between the upper and lower control limits Lucas (1985).

The CUSUM chart directly incorporated all the information in the sequence of sample value by plotting the cumulative sum of the sample values from a target value (Montgomery, 2009). Let \( \mu_0 \) be the target for the process mean, then, the CUSUM control chart is represented by

\[
C_t = \sum_{j=1}^{t} \bar{x}_j - \mu_0,
\]

where, \( C_t \) is the cumulative sum up to and including the \( t^{th} \) sample and \( \bar{x}_j \) is the average measurements for the \( j^{th} \) sample. Due to the combination of the information from several samples, CUSUM charts are more effective than Shewhart control charts for detecting small changes in the disease as well as for detecting small shift in the process mean.

The EWMA chart was introduced Roberts (1959). This control chart gives the most recent samples the greatest weight, and all previous samples weights decreasing in geometric progression form the most recent back to the first. Let the test statistic \( Z_t \) denote the geometric averages at time \( t \) while \( Z_0 = \bar{x}_0 \). The EWMA statistic is represented by

\[
Z_t = (1 - \lambda)Z_{t-1} + \lambda \bar{x}_t,
\]

\[
Z_t = \lambda \sum_{j=0}^{t-1} (1 - \lambda)^j x_{t-j} + (1 - \lambda)^t Z_0,
\]

where, \( 0 \leq \lambda \leq 1 \) and the weights \( \lambda(1 - \lambda)^j \) decreases geometrically with the age of the sample mean. And an out-of-control signal in these SQC chart is generated when the main statistic exceeds the control limit and when the signal is detected, an action should be taken to prevent the adverse effects from exceeding the target.
2.3.2 Traditional Multivariate SQC

It has also been noted that a characteristic may be determined by several (not necessarily independent) variables. Multivariate charts consider the mean vector and variance–covariance matrix of these variables. Several extensions of univariate control charts have been investigated: the Hotelling $T^2$ chart for individual observations (Hotelling, 1947; D. C. Montgomery & Klatt, 1972); the Multivariate CUSUM (MCUSUM), first proposed by Woodall and Ncube (1985) for bivariate variables, and generalized by Crosier (1988); and the Multivariate EWMA (MEWMA), first proposed by Lowry, Woodall, Champ, and Rigdon (1992) and combined with Principal Component Analysis (PCA) by Runger (1996), to improve its detection ability.

Hotelling’s $T^2$ control chart is an extended version of shewhart chart for monitoring multidimensional correlated process introduced by Hotelling (1947). The $T^2$ test statistic is optimal for testing the null hypothesis that the mean vector of a multivariate normal distribution is equal to a constant vector against the alternative hypothesis that the mean vector is not equal to the constant vector (Montgomery, 2009).

Suppose that the vector time series $\mathbf{x}_t = (x_{1t}, x_{2t}, \ldots, x_{pt})$ consists of $p$ univariate time series satisfying $\mathbf{X}_t \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$. Hotelling’s $T^2$ at time $t$ can be defined as,

$$T^2_t = (\mathbf{x}_t - \bar{\mathbf{x}})' S^{-1} (\mathbf{x}_t - \bar{\mathbf{x}}),$$

where, $\bar{\mathbf{x}}$ is an estimate of the mean and $S$ is an estimate of the covariance matrix. In retrospective Phase I application $t = 1, 2, \ldots, n$ and in prospective Phase II application, $t > n$. Typically the estimates of parameters $(\bar{\mathbf{x}}, S)$ are based on data from Phase I and the value of Hotelling’s $T^2$ can be compared with the upper control limit \( \frac{k(n+1)(n-1)}{n(n-k)} F_{1-a,k,n-k} \).
2.3.3 SQC Applied to Autocorrelated Data

Traditional SQC control charts, such as Shewhart, CUSUM, and EWMA charts, assume that the process observations are independent and identically distributed ($i.i.d$). Therefore, high false alarm rates often occur when the traditional SQC control charts are used to monitor highly autocorrelated processes (Alwan & Roberts, 1988; Harris & Ross, 1991). A common approach to monitoring auto-correlated observations is to first fit a time series model to the data, make a prediction about the future, and then monitor the residuals (observation–prediction), which will be a white-noise sequence if the process is in control. This approach was investigated by Reynolds and Lu (1997), who compared the performance of the EWMA chart on residuals and original observations.

As the extension of univariate framework, the multivariate time series modeling techniques can be used to remove the autocorrelation structure in observations collected from a multivariate process. Montgomery & Odonoghue (1999) and Hamilton (1994) introduced the multivariate time series model in vector ARMA (VARMA) form and related it to the state-space model. Noorossana and Vaghefi (2006) illustrated the use of the first-order vector AR(1) time-series model to decorrelate the multivariate process and the application of multivariate control chart to monitor the residuals.

Hotelling’s $T^2$-control chart is one of the earliest techniques in multivariate process control (Hotelling, 1947). It gives a general metric for turning measurement vectors into scalars that retain the essential information about whether the mean vector is indeed in control or not. A useful extension of this basic approach is discussed in (D.M. Hawkins & Olwell, 1998).
Healy (1987) derived a multivariate CUSUM (MCUSUM) procedure based on the sequential likelihood ratio test of multivariate variables on the scale of Hotelling’s $T^2$ statistic. Noorossana & Vaghefi (2006) applied the MCUSUM control chart to monitor the residuals from a vector AR(1) time series model. Lowry et al., (1992) presented a multivariate EWMA (MEWMA) control chart procedure as a logical extension of the univariate EWMA. Kramer and Schmid (1997) applied the MEWMA to the residuals from a vector AR(1) time series model.

The drawback of this approach is that a time-series model may not be readily available or that the model identified in Phase I does not adequately describe the process. Therefore, the process cannot be effectively decorrelated and false alarm rate is still higher than expected.

Other techniques for dealing with autocorrelated data are the Moving Centerline EWMA (MEWMA) and the model-free batch means control chart (Runger, 1996). Mastrangelo and Montgomery (1995) proposed a moving center-line EWMA (MCEWMA) approach to fit EWMA statistics to the observation values in order to minimize the one-step ahead prediction error, and thus combined the information of statistical control and process dynamics on a single control chart. They estimated the standard deviation of the one-step-ahead errors or model residuals by the mean absolute deviation (MAD). The MCEWMA approach is ideal for decorrelating IMA stochastic processes.

### 2.3.4 Phase I and Phase II

There are two phases for control charts application. The retrospective Phase I study is careful scrutiny of a sequence of past data. If based on a period when the process is deemed to be well-behaved, Phase I data provides a reference for how an in-control process behaves. Statistical
characteristics of the process are estimated in this phase are last used to determine control limit for Phase II. After the (regarded as) in-control process is studied and understood during Phase I, Phase II commences where the purpose is to monitor the process.

When applying SQC methods in actual monitoring fields, there are defining and separation concerns in Phase I and II phases. Most, if not all, control charting procedures assume known in-control parameters; otherwise two-phase monitoring is needed. Standard control chart usage involves Phase I and Phase II applications, which have two different and distinct objectives.

In Phase I, a set of process data is gathered and analyzed all at once in a retrospective analysis; trial control limits are constructed to determine if the process has been in control over the period of time during which the data were collected, and to see if reliable control limits can be established to monitor future production. This is typically the first step taken when control charts are applied to any process. Control charts in Phase I primarily assist operations personnel to bring the process under statistical control.

Phase II begins after a clean set of process data has been gathered under stable conditions and has been shown to be representative of in-control process performance. In Phase II, we use the control chart to monitor the process by comparing the statistics for each successive sample as it is drawn from the process to the control limits (Montgomery, 2009).

Shewhart control charts are highly recommended for Phase I; whereas CUSUM and EWMA charts have been shown to detect smaller process shifts in Phase II. It is generally acknowledged, though, that a combination of charts is likely to be advantageous.

In industrial quality control, it has been beneficial to carefully distinguish between the Phase I analysis of historical data and the Phase II monitoring stages. With phase I data, the focus is on
checking the stability of the process and estimating in-control parameter values for constructing Phase II methods. Phase I methods are usually evaluated by the overall probability of a signal; whereas run length performance is typically used for comparison purposes in Phase II, where the run length is the number of samples before a signal is given by the control chart. Steiner, Cook, Farewell, and Treasure (2000) used Phase I to establish a baseline performance for Phase II; however, in general, there is rarely a clear distinction between the two phases in health-related control charting.

It is recommended that a clear distinction be made in the health-related SQC literature between Phase I and Phase II applications and methods. However, it is often unrealistic to gather sufficient historical data for the in-control process, when monitoring clinical data for health monitoring purposes. Even if there are enough data to estimate the parameters, it cannot be guaranteed that the obtained parameters describe the in-control process. To deal with realistic limitations to define parameters during in-control process, Hawkins, Peihua, and Chang (2003) proposed change-point chart, whereby the need for Phase I parameter estimates is reduced and known parameters are not assumed.

### 2.4 Change-Point Chart

Change-point chart is a statistical technique used to detect and analyze changes in the pattern of a sequence of data. In the SQC area, the change-point model can be used to detect a change in the process parameters, such as mean and variance. Change-point has been used as an “add-on” procedure after a signal from an SQC chart has indicated that the parameters have changed after
the chart has displayed an out-of-control signal (Perry & Pignatiello, 2008). It simplifies further investigation to determine what the specific cause of the change is.

Change-point chart, however, can only identify the time of a normal process mean and variance shift. The basic univariate model (Hawkins et al., 2003) is

\[ X_i \sim \begin{cases} 
N(\mu_1, \sigma) & \text{if } i \leq \tau \\
N(\mu_2, \sigma) & \text{if } i > \tau,
\end{cases} \]

where, \( \tau \) is the change-point between the two segments of data. Both segments are independent and normally distributed with constant variance. For the univariate version, the change in mean is detected by sequentially computing the \( t \)-statistic to compare the mean of the “left” ( \( i \leq \tau \) ) and “right” ( \( i > \tau \) ) segments:

\[ T_{j,n} = \sqrt{\frac{(n-j)}{n}} \frac{\bar{X}_{j,n} - \bar{X}_{j,n}^*}{\hat{\sigma}_{j,n}}, \]

where,

- \( j \): Assumed change-point, \( 1 \leq j \leq n-1 \),
- \( n \): Possible range of observations, \( n = 3, 4, \ldots \)
- \( \bar{X}_{j,n} \): Average of the “left” segment,
- \( \bar{X}_{j,n}^* \): Average of the “right” segment, and
- \( \hat{\sigma}_{j,n} \): Pooled standard deviation of the two segments.

\( T_{j,n} \) are maximized across all possible \( j \) values, thereby obtaining the generalized likelihood ratio test statistics:

\[ T_{\text{max}}(n) = \max_{1 \leq j \leq n-1} |T_{j,n}|. \]
These statistics are then compared to a sequence of limits, \( h(j) \), which are obtained via simulation, while maintaining the conditional probability of a false alarm from any observation (given none present from previous observations) at a constant \( \alpha \) that satisfies the following probability equation (details on the simulation tabulated values of limits are contained in Hawkins, et al. 2003):

\[
\text{Probability } [T_{\text{max}}(n) > h(n) | T_{\text{max}}(n) \leq h(j), j < n] = \alpha.
\]

If \( T_{\text{max}}(n) > h(n) \), then the corresponding \( j \) that gives the \( T_{\text{max}}(n) \) is the change-point. In our application, some learning observations are necessary to obtain a credible detection signal.

Zamba and Hawkins (2006) provided the multivariate model and derivation of the \( T_{\text{max}} \) statistics:

\[
X_i \sim \begin{cases} 
N_p(\mu_1, \sigma) & \text{if } i \leq \tau \\
N_p(\mu_2, \sigma) & \text{if } i > \tau,
\end{cases}
\]

and

\[
\bar{X}_{j,m} = \frac{\sum_{i=j+1}^{m} x_i}{(m-j)}.
\]

The pooled covariance structure at \( j \) is:

\[
W_j = \frac{\left\{ \sum_{i=1}^{j} (x_i - \bar{x}_0,j)(x_i - \bar{x}_0,j)' + \sum_{i=j+1}^{n} (x_i - \bar{x}_{j,n})(x_i - \bar{x}_{j,n})' \right\}}{(n-2)},
\]

and its standardized difference between segments is:

\[
Y_{j,n} = \left[ \frac{j(n-j)}{n} \right]^{\frac{1}{2}} (\bar{x}_{0,j} - \bar{x}_{j,n}).
\]

The Hotelling \( T^2 \) statistics \( T_{j,n}^2 = Y_{j,n}' W_{j,n}^{-1} Y_{j,n} \) are computed to test differences between data at an assumed change-point \( j \) and are maximized over all possible values of
\[ T_{\text{max}}(n) = \max_{1 \leq j \leq n-1} |T_{j,n}^2|. \]

These statistics are compared to a sequence of control limits, \( h_{n,p,\alpha}(n) \). An intuitive property of the sequence would be to assign a constant probability of a false alarm for each \( n \) as the univariate case does. If the probability is a constant, \( \alpha \), then the sequence must satisfy the equation:

\[
\text{Probability}\left[ T_{\text{max},n} > h_{n,p,\alpha}(n) \mid T_{\text{max},n}^2 \leq h_{n,p,\alpha}(n); j < n \right] = \alpha.
\]

If \( T_{\text{max},n}^2 \) exceeds \( h_{n,p,\alpha}(n) \), then evidence exists that a shift has occurred and the corresponding \( j \), is the change-point. Since these limits cannot be derived mathematically, they are obtained via simulation (Zamba & Hawkins, 2006). Because the unknown-parameter change-point formulation requires no advance knowledge of the process parameters, it can be implemented from the initial stages of process monitoring. Testing is possible immediately.

To lessen the possibility of false alarms, some learning observations are necessary. Here, an initial set of twenty-five learning observations was assumed. The closed form approximation for \( \alpha = 0.002 \) (Zamba & Hawkins, 2006) is:

\[
\ln(h_{n,p,0.002}) = 0.043 + 0.221p + \frac{p+4}{50} \ln(n - 25).
\]

To consider one-sided detection, the choice for specificity is 99.8% (\( \alpha = 0.002 \)), which is equivalent to one false alarm every 400 weeks, or roughly once every eight years.
CHAPTER 3

A SYNDROMIC SURVEILLANCE SYSTEM

FOR CLINICAL AND NON-CLINICAL HEALTH DATA

Syndromic surveillance is a type of public health surveillance with the specific goal of detecting disease outbreaks earlier than conventional reporting of confirmed cases. Henning (2003) demonstrated the importance of detecting disease outbreaks earlier than traditional approaches by studying already-confirmed cases and analyzing outbreaks in terms of size, spread, and rate. His work showed that by monitoring disease trends, it is possible to predict when outbreaks will occur. In this chapter, we propose a health monitoring framework that provides a customized to syndromic surveillance in the event that clinical and non-clinical data can leveraged for analysis.

3.1 Introduction

The term “syndromic” refers to the respiratory, gastrointestinal, rash, neurologic, and sepsis syndromes, which are defined based on ICD-9CM (International Classification of Disease, 9th revision Chief Complaints: http://www.cdc.gov/nchs/about/ major/dvs/icd9des.htm). If the initial symptoms categorized in ICD-9CM are observed, it is definitely possible to detect disease outbreaks earlier than traditional disease detection methods that depend solely on the clinical or laboratory confirmation of a particular disease. In line with the goals of earlier detection protocols, current research focuses on gathering information about patients’ symptoms (e.g., cough, fever, shortness of breath) during the early symptom (prodrome) period of illness.
The epidemic curve for persons with earliest symptom onset and the epidemic curve for those with severe symptoms of an illness can be depicted graphically, as shown in Figure 3.1. The time \( t \) can be represented as the difference between symptom onset for an increasing number of cases caused by infectious disease and subsequent patient visits to a healthcare facility resulting in a definitive diagnosis. The goal of early detection corresponds to the national influenza pandemic preparedness action plan developed by the CDC, the goal of which is to limit the burden of the disease, minimize social disruption, and reduce economic losses attributable to it (Striaks et al., 2002). In addition, early detection enabled by sensitive surveillance enables intervention efforts to take place as soon as possible so that they will be as effective as possible; such efforts could include quarantine, prophylactic use of medication, and heightened community awareness.

Figure 3.1. Syndromic surveillance helps prevent severe illness by treating patients with initial symptoms earlier than traditional detection. Here \( t \) is the time between detection by syndromic surveillance and detection by traditional methods.

Prior to the advent of syndromic surveillance, public health surveillance had generally been limited to retrospectively analyzing medical data. However, the retrospective analyses often proved to be too time-consuming, failing to signal an alert within a reasonable time period.
Timely alarms may help prevent spread of the disease and has the potential to significantly improve effective response to an outbreak or bioterrorism attack, particularly if the disease is contagious. The main goal when using these methods is to detect outbreaks as quickly as possible, in order to optimize the response time for public health officials.

In an effort to reduce the delay time, public health organizations have begun to implement syndromic surveillance systems that are intended to provide for earlier detection. Syndromic surveillance systems have been developed using SQC. Further, the notion of abnormality or aberration detection in health surveillance is similar to the idea of detecting an out-of-control condition in SQC (Bueher et al., 2004). After something unusual is notified in health data patterns, an epidemiologic investigation will have further studies with a view to confirming that a disease is dispersing, similar to a search for assignable causes pursuant to an SQC control chart signal. Because of the goal of detecting an abrupt increase in health data for syndromic surveillance purposes, monitoring typically relies on a one-sided control chart or surveillance method in public health (Fraker et al., 2007).

In the past few decades, a number of methods in SQC have been developed for the detection of changes in populations (Montgomery, 2009). These methods typically estimate change points in a sequence of disease events, a time series of population rates, or the determination of the application of control limits to the behavior of a system. In health surveillance, some simple SQC methods, such as CUSUM and EWMA, are available to assist in the assessment of change or in-control behavior (Woodall, 2006).
Figure 3.2. ILI incidence (source: French Sentinel Network, www.sentiweb.org) and several classes of drugs likely to be purchased or prescribed in ILI context (source: IMS France).

However, control charts in SQC require that the observations are independently and normally distributed. Monitoring the data that violate these general assumptions yields uncertain results (Shmueli & Fienberg, 2005; Burkom et al., 2006). Unlike industrial processes with a clear specification or target value and a more controlled environment, public health data follow natural epidemiologic processes. As shown in Figure 3.2, the time series of the weekly number of patients with ILI per 100,000 persons in France (bold solid line) is a non-stationary process with a high level of autocorrelation. Therefore, the nature of the health data emphasizes a barrier to the direct use of traditional SQC methods either based on univariate or multivariate monitoring tools for syndromic surveillance.

This study proposes a multivariate monitoring methodology based on the change-point method applied to a public health surveillance system. Our objectives are to: (1) precondition health data in order to render possible their monitoring by SQC tools; and (2) apply this proposed
approach to multivariate monitoring of ILI and OTC and prescribed medication sales data in France in the context of influenza surveillance in order to explore whether it provides earlier signal.

The proposed monitoring scheme first consists in modeling the baseline activity of ILI during the non-epidemic season (distinguished from the epidemic season using the Serfling’s regression) based on the method of analogues. These forecasted non-epidemic values are then subtracted from observed data and resulting residuals are passed through the ARMA filter before being analyzed with the change-point model.

This monitoring methodology is applied to multiple data streams of ILI (times series of incidence in 22 regions) and drug sales data (several drug classes) in order to explore which of the two data sets provides the earliest signal of outbreaks with the reference to the univariate monitoring of clinical ILI incidence (national level). As shown in Figure 3.2, the drug sales data follow similar patterns as the ILI data, but the drug sales data are shifted earlier in time. The trend of the ILI curve lagging the drug sales curve may imply possibilities for early warning for outbreaks.

3.2 Literature Review

Traditional public health surveillance is retrospective in nature and mainly relies on physician and laboratory reporting and manual analysis of surveillance data. Yet, these techniques have limited capability for delivering timely detection of a disease’s outbreak (Tsui et al., 2003). However, recent research in health surveillance systems complements traditional public health
monitoring with routine data analysis in order to create and send earlier warnings of potential threats.

Current syndromic surveillance systems run multiple simultaneous univariate SQC procedures, each focused on detecting an increase in a single dimension. Woodall and Ncube (1985) first proposed the application of simultaneous univariate CUSUMs in a multivariate application. Multiple simultaneous univariate procedures have the advantage of being easy to implement and interpret, but they can be less sensitive than multivariate methods to some types of changes. However, unless the signal thresholds of the multiple simultaneous procedures are properly set, they can suffer from a higher than optimal combined false alarm rate.

Rogerson and Yamada (2004) evaluated multiple univariate CUSUMs versus a directionally invariant multivariate CUSUM for monitoring changes in spatial patterns of disease. Testik and Runger (2006) developed directional multivariate procedures and compared multivariate procedures that look for an increase in a pre-specified subset of a mean vector’s components. Their procedures also allowed the remaining components to either increase or decrease as well as allowing for a shift of the mean vector in the direction of a specific vector and/or a shift in the mean vector corresponding to an increase in one or more components of the mean vector, while existing works on evaluating the effects at least one of the components of the mean vector has increased.

Parametric methods have also been investigated to model disease incidence data and to provide alarm thresholds. Trigonometric regression was first introduced to model non-epidemic pneumonia- and influenza-related deaths (Serfling, 1963). This method has also been used to model the baseline activity of influenza-like syndromes in France, and therefore, to derive
weekly ILI epidemic thresholds (Costagliola et al., 1991). The main constituents of the regression are the sinusoid and linear functions that have a fixed cycle length and amplitude by nature. Thus, the mathematical form can allow neither a variation in the yearly epidemic length nor variations in the height of the peak. Generally speaking, the problem lies in identifying observations belonging to non-epidemic periods when estimating a threshold from non-epidemic data.

Poisson regression has also been used for epidemiological time series since the data are counts-based (Hu et al., 2007; Kuhn et al., 1994; Tobias et al., 2001; Zeger, 1988). However, these studies do not capture the autocorrelation inherent in infectious disease incidence. The autocorrelation was considered by including an autoregressive term with time series (Vergu et al., 2006). Time series methods (Bowie & Prothero, 1981; Campbell, 1994; Helfenstein, 1986; Hu et al., 2004; Hu et al., 2007; Hussain et al., 2005; Quenel & Dab, 1988; Stroup et al., 1988; Tobias et al., 2001) are able to handle autocorrelation successfully, but start to fail when the normality assumption is not satisfied. Morton and Finkenstädt (2005) used a Monte Carlo Markov Chain (MCMC) approach to model disease incidence time series, and Le Strat and Carrat (1999) used a hidden Markov model that allows for switching between epidemic and non-epidemic states.

Working on Swedish influenza data, Anderssson et al. (2007) noticed that the data differ in both height and growth from year to year, and there is no consistency in the up or down phase. In the beginning, they suggested using Poisson distribution for approximation near the peak. But later, they noticed that Gaussian distribution was more reasonable than Poisson distribution for the approximation. As shown in Chandra et al. (2007), to obtain a satisfactory fit on the ILI data
in the United States, different arbitrary segments of the year must be fitted separately to achieve
different parameterization. However, the inconsistency of the timing and height of the peak
makes results obtained in this way difficult to use for framing predictions. Anderssson et al.
(2007) further suggest using the nonparametric model and the monotonous property of a rise in
the incidence instead of using some parametric methods, as there is no strong evidence to support
a certain distribution.

In addition to the use of appropriate statistical methods, a faster detection of outbreaks could
be achieved by the joined use of multiple data streams (Vergu et al. 2006, Rolka et al. 2007).
Because a single data source does not capture all the individual behaviors in the outbreak, the
statistical power to detect an outbreak with a combined data set can be improved (Kulldorff et al.
2007). Recently, non-clinical data such as sales of over-the-counter (OTC) and prescribed
medication and healthcare products, absenteeism from work/school, calls to nurse hotlines, can
also be collected. Since these non-clinical data represent behavior of sick people before a visit to
the physician, they can be used as an early indicator of disease outbreaks. Modern public health
surveillance, utilizes multiple data streams from both clinical and non-clinical sources, to
identify disease clusters earlier than would laboratory-confirmed cases.

There is some discussion about the need to apply non-clinical data to syndromic surveillance.
Alternative data sources can provide useful data about behavior patterns; for example,
school/work absenteeism and drug sales data could be valuable in understanding an outbreak of a
disease. Thus, non-clinical data certainly has the potential to provide more sensitive detection
ability in syndromic surveillance because non-clinical data reflect the behavior of sick people
before they visit a doctor; that is, they provide data on behaviors that are precursors to the diagnosis of human disease.

3.3 Method

Many monitoring systems in syndromic surveillance use control charts originally intended for industry applications (Lawson & Kleinman 2005; Griffin et al. 2009). The challenge in applying standard SQC methods is that data used in syndromic surveillance generally violates classic SQC assumptions. Another practical issue is that many officials are having problems determining the true base line period against which the current data should be compared in order to check for true anomalies versus false positives. This is similar to the Phase I problem in SQC, in which an in-control process parameter estimation needs to be completed in order to establish the control charts used for actual monitoring in Phase II. Without the separation between Phase I and Phase II, an appropriate approach is the change-point model that can monitor for the shift in mean and variance of data with unknown initial parameters through sequential analysis (Hawkins & Zamba, 2005). Compared to other traditional monitoring tools such as Shewhart, CUSUM, or EWMA charts, change-point has been used as an “add-on” procedure after a signal from an SQC chart, which indicates the time when the parameters change after the chart displays an out-of-control signal (Perry & Pignatiello 2007; Perry et al. 2006; Pignatiello & Samuel 2001). It simplifies further investigation to search for the special cause of the change.
In order to transform the initial data for a monitoring based on the change-point method, we propose a surveillance methodology illustrated in Figure 3.3. First, the raw data (observed data), $y$, are thresholded ($\hat{y}_{\text{thresh}} < \text{threshold}$) after determining the threshold by Serflings’s trigonometric regression (1963), in order to differentiate between non-epidemic (baseline cases) and epidemic values (abnormal cases). The nonparametric method of analogues (Viboud et al. 2003) is used to build forecasted values of the non-epidemic period ($\hat{y}_{an}$). The residuals $r_{an}(= \hat{y}_{\text{thresh}} - \hat{y}_{an})$ are used to build an autoregressive and moving average (ARMA) filter. Then, the differences $r(= y - \hat{y}_{an})$ between observed and predicted data under the non-epidemic hypothesis are passed through the ARMA filter to remove the autocorrelation. Finally, a change-point method is applied to filtered residuals ($r$). During non-epidemic periods, residuals ($r$) will be independent and normally distributed, which satisfies the assumptions of
change-point chart. However, when observed values \((y)\) significantly exceed the forecasted baseline level \((\hat{y}_{an})\), the residuals will increase significantly. Therefore, the change-point monitoring detects the condition and signals the start of an epidemic period.

In the following, we detail each of the stages of this surveillance methodology. For sake of simplicity we present these different methods when applied to ILI univariate data, but the procedure detailed in Figure 3.3 was applied independently to univariate and multivariate ILI incidence data and to multivariate drug sales data.

### 3.4 Data

All the data sets used in this study encompass the same range: between week 36 of 2000 to week 8 of 2009. A first set of data (from week 36 of 2000 to week 10 of 2007) was used to build predictions for the non-epidemic season, while the remaining data (from week 11 of 2007 to week 8 of 2009) were used to simulate real-time monitoring.

### 3.4.1 Drug Sales Data

The database for drug sales, freely provided by IMS France (www.imshealth.com), consists of weekly prescription and OTC drug sales. These data come from 6,000 to more than 12,000 retail pharmacies (over 50% of all French pharmacies, not including hospital pharmacies). Experts from the World Health Organization (WHO) Collaborating Center pre-selected 20 classes of medications likely to be prescribed or purchased for ILI. Descriptions of the drugs appear in Table 3.1, where each entry represents a family of drugs.
Table 3.1. Twenty classes of drug sales likely to be purchased or prescribed in an ILI context.

<table>
<thead>
<tr>
<th>Vitamin C only</th>
<th>Anti-inflammatory rhinological drugs without corticoids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tetracycline in association with another medication</td>
<td>Anti-allergic rhino-preparation</td>
</tr>
<tr>
<td>Penicillin</td>
<td>Nasal decongestant</td>
</tr>
<tr>
<td>Cephalosporin</td>
<td>Topical rhino-preparation</td>
</tr>
<tr>
<td>Macrolides</td>
<td>General rhino-preparation</td>
</tr>
<tr>
<td>Other antibiotics</td>
<td>Anti-inflammatory decongestant, pharynx</td>
</tr>
<tr>
<td>Antirheumatics non-steroidal only</td>
<td>Expectorants</td>
</tr>
<tr>
<td>Myorelaxant with peripheral activity</td>
<td>Cough suppressant only</td>
</tr>
<tr>
<td>Analgesic nonnarcotic and antipyretic</td>
<td>Cough suppressant in association with another medication</td>
</tr>
<tr>
<td>Rhinocorticoids without anti-inflammatory agent</td>
<td>Cough suppressant with bronchial-pulmonary antibiotic</td>
</tr>
</tbody>
</table>

3.4.2 Influenza-Like-Illness (ILI) Data

Data on ILI incidence arises from the French Sentinel Network (FSN), a computerized public health surveillance system. Since November 1985, this system has collected reports from 1,260 general practitioners who are voluntary participants throughout France. Their participation requires logging onto the FSN at least once a week and updating a web-accessible database with information on communicable diseases including ILI. Ongoing, weekly national and regional ILI incidences are freely available on the web (http://www.sentiweb.org). In reference to a standardized surveillance practice, the participating general practitioners enter an individual report for each visit involving ILI, including symptoms such as sudden fever of more than 39°C, myalgia, and respiratory symptoms. From the raw reports, a set of routine procedures generates ILI incidences every week for each administrative area (22 regions in France). This study used
the series of weekly ILI incidences (number of cases per 100,000 populations) for France and its 22 administrative regions.

3.5 Result

The focus is on the monitoring effort to detect the onset of a yearly influenza epidemic from week 11 of 2007 to week 10 of 2009. Univariate change-point monitoring of normalized ILI (national level) residuals and multivariate change-point monitoring of normalized ILI residuals at the regional level (22 regions) and of 20 classes of normalized drug sales residuals were performed.

![Graphs showing ILI incidence comparison](image)

(a) National ILI incidence: 2007.11\(^{\text{th}}\) week ~2008.10\(^{\text{th}}\) week
(b) National ILI incidence: 2008.11\(^{\text{th}}\) week ~2009.10\(^{\text{th}}\) week

Figure 3.4. French Sentinel Network (www.sentiweb.com) Results

The residuals monitored represent the difference between the observed raw data and their respective forecasts obtained using the method of analogues. Detecting an aberration in residuals could indicate that the observed value is significantly different (higher) than the predicted baseline value and serves as a warning of an upcoming epidemic period. These warnings are compared to the onset of an epidemic season as defined by the FSN (the first week when the
national ILI incidence exceeds a baseline non-epidemic threshold given by the upper limit of the 95% confidence interval of the Serfling model (Costagliola et al. 1991). Figure 3.4 indicates that the onsets of epidemic season which FSN officially announce based on the epidemic threshold of Serfling model were defined as 54th week in 2007 and 51th week in 2008.

The thresholded data ($\hat{y}_{\text{thresh}}$) can be obtained by truncation when the observed data ($y$) is larger than the epidemic threshold. The thresholded data ($\hat{y}_{\text{thresh}}$) are used for the non-parametric method of analogues which creates the analogue forecast ($\hat{y}_{\text{an}}$).

![Figure 3.4](image1.png)

(a) ILI incidence: national  (b) ILI incidence: 15th region  (c) Drug sales: cough suppressant

Figure 3.5. Comparison between observed data, non-epidemic thresholded data and forecasted values obtained using the method of analogues.

Figure 3.5 shows that the method of analogues provides a good fit with a 7% absolute mean error on average across all series. But the residuals between $\hat{y}_{\text{thresh}}$ and $\hat{y}_{\text{an}}$ still do not satisfy the conditions for using change-point chart as illustrated in Figure 3.6.
The residuals between thresholded data and analogue forecast appear highly autocorrelated. Thus, a simple ARMA model is used to transforms residuals to normally i.i.d. errors. Normality and independence assumptions are finally satisfied after filtering as represented in Figure 3.7, as these assumptions are required for applying the change-point model.

3.5.1 Univariate Change-Point Chart (ILI incidence at national level)

Univariate change-point monitoring was performed on normalized ILI residuals. The test statistics, $T_{max}(n)$, were updated as new observations become available and the $T_{max}(n)$ data
were compared to the limit, \( h(n) \), as simulated in Hawkins et al. (2003) with specificity of 99.8%. Figure 3.8 shows the change-point monitoring for ILI. Both monitoring efforts began with week 11.

The outcomes obtained by univariate monitoring provided earlier warning than the official start of the epidemic season as defined by the FSN. As represented in Figure 3.8, the univariate change-point model suggested outbreaks 3 weeks earlier in 2007 and 9 weeks earlier in 2008, as compared to FSN definitions of the start of an epidemic period.

![Figure 3.8](image.png)

Figure 3.8. Univariate change-point applied to transformed ILI incidence (national) to detect epidemic onset. (ILI incidence data: solid line, change-point threshold: dashed line)

### 3.5.2 Multivariate Change-Point Chart (ILI incidence in 22 regions and 20 classes of drug sales)

Using the ILI data, the outcomes of the multivariate monitoring showed overall good agreement with those provided in the univariate change-point model. As shown in the Figure 3.9, the multivariate change-point chart based on regional ILI incidences (22 regions) gives an earlier
indication of the outbreak than the univariate method: 3 weeks earlier in 2007 and 2 weeks earlier in 2008. This corresponds to the detection of a signal 6 and 11 weeks respectively before the epidemic onset. Earlier warning about the onset of the epidemic season can come from the ability to detect irregularities not only in every variable but also in abnormal relationships among variables by considering multivariate monitoring.

Figure 3.9. Multivariate change-point applied to transformed ILI incidence (22 regions) to detect epidemic onset. (ILI incidence data: solid line, change-point threshold: dashed line)

(a) Regional ILI incidence: 2007 11th week ~2008.10th week
(b) Regional ILI incidence: 2008 11th week ~2009.10th week

Figure 3.10 illustrates that a multivariate monitoring of drug sales data through a change-point chart provides even earlier warnings than the multivariate monitoring of ILI data: a signal is provided 13 weeks earlier in 2007 and 14 weeks earlier in 2008 than the onset of the epidemic season as defined by the FSN.
3.5.3 Discussion

The goal of syndromic surveillance is to track health-related data in such a way that early warnings of disease outbreaks may be obtained without having to wait for laboratory confirmed diagnoses. Such warnings not only enable patients who are exhibiting prodromal symptoms to be treated with specialized care, but they also mobilize a rapid response and thereby reduce morbidity and mortality (WHO, 2009).

This chapter illustrates the application of an SQC monitoring method for public health surveillance. Based on the principles of SQC, an aberration is suspected when the number of reported cases (or drugs sold) exceeds the expected level derived from historical data.

The data have atypical statistical characteristics, and it is nearly impossible to know which parameters ensure process stability. To solve these problems, we proposed an alternative approach consisting of pretreatment steps using preconditioning (trigonometric regression) and a
forecasting method (the method of analogs) to transform the data to fit a general SQC approach. The change-point method, a SQC tool that does not require knowledge of in-control parameters, was used as a final monitoring step. By calculating a trigonometric threshold and applying the analog methodology as preconditioning steps, observations could be predicted for a non-epidemic season, creating the baseline data required for the change-point method. The residuals (i.e., differences between observed and predicted values) could then be monitored using change-point methodology to determine if the number of reported cases (or drugs sold) was significantly higher than the expected value.

Monitored results can largely depend upon the chosen threshold value from the Serfling regression model, since data assumed to be representing epidemic periods are deleted when the epidemic threshold value is calculated. In this study, the epidemic season was assumed to range from the 49th week to the 12th week of the following year, the weeks commonly observed as the epidemic season for every ILI (national and regional) and drug sales time series. Ideally, applying SQC to different epidemic seasons retrospectively would produce different epidemic thresholds. Additional investigation would be useful to quantify the effects of threshold values on the performance of the monitoring scheme.

The method of analogs is a useful tool for providing predictive value when comparing other time series as long as the stationary condition is satisfied. The method is a non-parametric approach that makes no assumptions about the distribution with respect to specific mean or variance levels, so it may have better outcomes than other time series models. However, the data in syndromic surveillance applications tend to be non-stationary with big peaks. This chapter explored integrated data modeling with a baseline activity of disease based on an analog forecast
using the threshold values obtained by trigonometric regression as the upper boundary. This integrated modeling approach was proven to be appropriate even for forecasting drug sales data and ILI data, which violate the stationary characteristic because the correlations between analog forecasts and threshold data were above 0.75 and as high as 0.9.

The pretreatment we propose on epidemic data that is non-stationary, non-normal and highly autocorrelated enables the use of change-point, which usually requires data to be independent and identically distributed. The change-point method is used to find the point in time where an underlying series of observations changes by reapplying the change-point test to all accumulated data when a new observation is added. Calculating these statistics enables parameters of in-control processes to be estimated. This procedure is repeated in such a way that the probability of a false alarm remains constant. The false alarm rate determines the control limit, which is the threshold between in-control and out-of-control states. The tradeoff between the cost of false alarms and the cost of delaying an infectious disease outbreak warning must be considered for each specific surveillance scenario.

The change-point method outlined in this chapter is used to evaluate how the sustained mean increases with multiple data in accord with the goal of syndromic surveillance: finding continual increases in observed indication, and ultimately detecting infectious disease outbreaks. Additional monitoring methods that can consider changes in a covariance matrix (Zamba and Hawkins, 2009) may be able to observe abnormal signals earlier with pretreated data.

The change-point method provides information as to when the onset of an epidemic season actually occurs. This could help public health decision makers determine the correct interventions and when to implement them (e.g., vaccinations, which are effective before
widespread outbreaks). It may also assist in planning to better handle the burden on resources that disease outbreaks create. Further research is required to fully understand the use of the change-point method in these type of monitoring applications.

The surveillance scheme proposed in this study was applied to ILI incidence (at both national and regional levels) and drug sales related to ILI symptoms. We showed that monitoring with a multivariate change-point procedure detects outbreaks a few weeks earlier than univariate change-point monitoring on the same ILI data. This difference is due to the fact that multivariate methods consider the relationships between variables. If variables are independent, applying a univariate control chart to each individual variable is a possible solution. However, most data sets in the study are related to each other, so considering them separately may not sufficiently account for the dependence between variables, especially in terms of the spatial autocorrelation of regional data in the ILI data set.

Moreover, medication sales, as non-clinical data, provided an indirect but earlier signal for an influenza outbreak because non-specific symptoms in the prodromal phase of many diseases may be self-treated before visiting a doctor. This result is empirically supported by visual examination of the data. The ILI data set has a long string of low values in non-epidemic seasons. The strings then steadily rise to peak values and decrease gradually in a similar manner. For drug sales, comparable patterns with larger peaks preceded by abrupt increases can be observed. Therefore, drug sales data can provide an early signal announcing a coming increase in the number of ILI cases.

This monitoring approach based on clinical and non-clinical data is very promising since it illustrates the properties of efficient warning indicators that could be used by public health
authorities. However, further analyses will have to be performed in order to refine the interpretation of such early detection of epidemics. Indeed, the prediction of an ILI outbreak several weeks in advance (based on the correlation between ILI cases and classes of drugs used in ILI contexts) could be the result of a complex interplay between different factors or involve other indirect mechanisms.

For instance, selected medication classes could be direct indicators that ILI cases have a more sensitive threshold than the target signal itself: their use (and hence, related sales revenue) begins to grow resulting from an increase in ILI activity even earlier than this increase could be identified from an ILI time series. Conversely, other pathologies synchronized with ILI activity might influence the sales pattern of chosen medication classes, which can cause it to behave similarly to indirect predictors of ILI outbreaks.

More refined analyses should also be carried out at regional or sub-regional levels in order to detect potential unusual local events involved in signal detection. Drug consumption for ILI symptoms varies from region to region. When external factors such as infectious disease or bioterrorist attacks occur in one specific region, a change in medication consumption occurs in that region. The change will eventually impact other regions and the consumption of drugs in those regions will begin to rise as well. Thus, the change in consumption may provide an advanced signal for outbreaks even if the number of variables considered is too large and the limitation for specifying the multi-regional covariance structure is met. Additionally, simulation study could be conducted to further explore performance of the methodology as well as vis-à-vis other methods such as a scan statistic.
In conclusion, the syndromic surveillance approach described here uses a series of data modeling and monitoring processes that combine preconditioning (trigonometric regression), a forecasting method (the method of analogs), and a monitoring method (the change-point) to detect outbreak points earlier than conventional ways by adapting to the statistical characteristics in health data. This method estimates temporal changes in disease dispersion based on previously observed surveillance data, including both clinical data (ILI) and non-clinical data (drug sales data). Throughout the three comparison trials (single clinical data, multiple clinical data sets, and multiple non-clinical data sets) in this chapter, multiple data sets from non-clinical sources proved to produce the earliest signal for outbreaks. The results in this study confirm that drug sales data could not only be used as an independent additional source of information to predict outbreaks, but also provide an earlier signal than ILI data used alone. The proposed method is both practical and relatively simple to implement, even with multiple streams of data that are interdependent. The proposed data modeling and monitoring scheme has many possible applications for other communicable and non-communicable diseases, not just for influenza surveillance, as discussed above. Detecting onset moments for hypertension or diabetes - the main causes of death in the USA (WHO, 2009) - could be a potential application area for this method.
CHAPTER 4

UNKNOWN PARAMETER MULTIVARIATE STATISTICAL QUALITY CONTROL MODELS TO MONITOR CHRONIC DISEASE

Healthcare monitoring in practical applications often demands that two or more quality characteristics be tracked. Traditional multivariate statistical quality control charts focus mainly on mean shifts, even though the process mean and variability change during the monitoring period. Furthermore, most suggested traditional SQC tools assume that in-control parameters are known, when in reality, process parameters are not given a priori and insufficient serial measurements exist to estimate in-control parameters. In light of the challenges faced in healthcare monitoring, we consider the implementation and performance of the multivariate change-point (MCP) chart and multivariate self-starting exponentially weighted likelihood ratio (MSS-EWLR) chart.

4.1 Introduction

Most healthcare monitoring guidelines specify a particular target that should be achieved, but this ignores the uncertainty and imprecision involved in making clinical measurements. Even if the true underlying value of a disease marker (such as cholesterol or blood pressure) in an individual patient is stable, we would expect to observe variability in the measurement over time. To deal with this issue, it is appropriate to adopt statistical quality control (SQC) methods, which take into account inherent variability in healthcare monitoring.
SQC is becoming widely accepted in the health domain as a technique for monitoring processes and outcomes (Grigg, Farewell, & Spiegelhalter, 2003; Sherlaw-Johnson, Gallivan, Treasure, & Nashef, 2004; Spiegelhalter, Grigg, Kinsman, & Treasure, 2003). Although most SQC applications are routinely used in industrial settings to monitor process behavior and performance, SQC has been adopted to monitor the health of individual patients with chronic diseases (Winkel & Zhang, 2007; Woodall, 2006).

An SQC chart is a chronological graph of process data with statistically defined control limits. The main objective of SQC is to identify unusual patterns of variation and evidence of special causes when process values fall outside the limits. In clinical settings, we can prospectively use control chart procedures to determine whether expected changes in the physiological variable structure occurred (indicating improvements in process of care) and to explore the underlying reasons for observed patterns. Also, SQC tools can also be used in health monitoring to detect unexpected signals for special cause variations, such as disease onset. They have been applied in clinical settings in order to identify when biological characteristics change, thus increasing or decreasing health risk.

To date, attention has focused mainly on potential applications of SQC tools in healthcare monitoring. However, most control charts described in previous work tend to introduce simpler applications by favoring univariate over multivariate monitoring, and by monitoring only location shifts rather than simultaneously monitoring location and scatter. In actual healthcare monitoring applications, there is a need to track several quality characteristics, which are generally correlated, simultaneously. Monitoring clinical outcomes requires detecting subtle to moderate changes in a patient’s underlying conditions and observing those changes in both the
mean vector and covariance matrix. Hence, it is necessary to use a more sensitive multivariate charting process that can detect changes in both parameters (instead of depending only on the mean) in order to highlight correlations between quality characteristics.

Furthermore, general charting processes assume that in-control parameters, which are necessary to define control limits, are known. In actual practice, in-control parameters are rarely known. To deal with this limitation in applications, most practitioners use the “plug-in” method where a large data sequence is gathered while the process is believed to be in control. This technique can be acceptable in controlled industrial environments with clearly defined specifications or targets. However, when monitoring an individual patient, it is often unrealistic to gather sufficient historical data for the in-control process. Even if there are enough data to estimate the parameters, it cannot be guaranteed that the obtained parameters describe the in-control process.

In this chapter, we survey several multivariate charting methods that enable mean and variability to be monitored in a single chart, resulting in more sensitive monitoring. Based on this review, we investigate the multivariate self-starting control chart combined with the exponentially weighted likelihood ratio (MSS-EWLR) chart and the multivariate change point (MCP) chart. These approaches have the added benefit of requiring no distinction between in-control processes and monitoring processes which may greatly aid the ability to monitor processes in healthcare.

The remainder of the chapter is organized as follows. In the next section, we provide the general multivariate framework for simultaneous monitoring and for monitoring without in-control parameter information, and explain our proposition for developing these control charts
for the healthcare environment. Following this, we outline the methodology of a simulation study we used to investigate how each chart performs for various correlation structures, patterns of mean, or covariance shifts. We illustrate the application of our proposed methods using real ambulatory blood pressure data. In the final section, we discuss and make several recommendations regarding the implementation of SQC charts in chronic disease management, and data management when the basic assumptions of SQC are violated.

4.2 Literature Review

4.2.1 A General Framework for Simultaneous Monitoring

There are many clinical monitoring settings where simultaneous monitoring of two or more quality characteristics is necessary, i.e., there are \( p \) variables given by \( x_1, x_2, \ldots, x_p \), arranged in a \( p \)-component vector \( x = [x_1, x_2, \ldots, x_p]' \). The most widely used basis for monitoring these variables is the multivariate normal distribution. In this setting, it is assumed that while in control (IC), the quality characteristics follow independent multivariate normal distribution with some mean vector \( \mu \) and covariance matrix \( \Sigma \). However, these quality characteristics are usually correlated, and the process mean and variability may change simultaneously during the monitoring period from \((\mu_0, \Sigma_0)\) to \((\mu_1, \Sigma_1)\). Especially in multivariate settings, changes in the process mean vector can be masked by unsuspected changes in the covariance structure based on a sudden change in the correlation among quality characteristics (Zamba & Hawkins, 2009). Therefore, it is desirable to construct a control chart that can not only detect changes in the process mean, but also in process variability.
Most traditional control charts (e.g., Shewhart, Hotelling $T^2$, MCUSUM, MEWMA) for variable data focus on developing separate procedures for monitoring process location and spread. In order to monitor process shifts, two control charts – one for monitoring process mean and the other for monitoring process variability – are run concurrently. This practice can be inefficient.

In recent years, scholars have attempted to develop control charts using only one plotting chart to monitor changes in both the mean and covariance matrix. The single chart is created by combining process parameter statistics (i.e., mean and variability). The transformed statistics for signal control charts are then plotted against time. When there is a change in a process, the single chart should be able to issue an out-of-control (OOC) signal. The effectiveness of the control charts is evaluated based on its average run length (ARL). In general, the IC ARL ($ARL_0$) should be long, while the OOC ARL ($ARL_1$) should be relatively short.

In general, there are three main approaches to developing a control chart that simultaneously monitors process mean and variability. One is to combine two control charts designed for monitoring the mean and dispersion. The development of combined single control chart originates from work by (Alt, 1984) based on the traditional combination of the Hotelling $T^2$ and generalized variance $|S|$ control charts. (Reynolds, Marion, Cho, & Gyo-Young, 2006) proposed a MEWMA chart based on the sample means and the sum of squared deviations from target. Additionally, Hawkins and Maboudou-Tchao (2008) considered a combination of the multivariate and exponentially weighted moving covariance matrix charts.

Another method to single simultaneous monitoring is the multivariate maximum chart. The basic idea behind multivariate maximum charts is to transform the monitoring statistics for mean
and covariance to standardized normal random variables by taking inverse function of the corresponding distribution. The main statistics for multivariate maximum charts can be defined as the maximum absolute values of charting statistics in terms of mean vector and covariance matrix. The maximum values of these standard normal readings are then applied in a traditional multivariate charting procedure. Proposed multivariate maximum charts include the Multivariate Maximum Control Chart (Thaga & Gabaitiri, 2006), the Multivariate Maximum EWMA chart (Chen, Cheng, & Xie, 2005), and the Multivariate Maximum CUSUM control chart (Cheng & Thaga, 2005).

The final simultaneous monitoring method uses generalized likelihood ratio (GLR) test statistics to implement a single control scheme for detecting shifts in both the mean and the covariance matrix. The GLR test is a common and effective statistical approach for testing composite hypotheses with regard to shifts in the mean and/or covariance matrix. Multivariate simultaneous charting processes based on the GLR test include the multivariate change-point method (Zamba & Hawkins, 2006, 2009) and the multivariate exponentially weighted likelihood ratio chart (Zhang, Li, & Wang, 2010). Considering the inherent benefits of a classical GLR test, these charts are quite robust and sensitive to various types of shifts.

4.2.2 A General Framework for Unknown Parameter Multivariate Control Charts

One major disadvantage to applying existing multivariate control charts is that they require the parameters to be known a priori. In general, traditional SQC charts rely on known or assumed true in-control parameters from past data sequences (Phase I). Parameter information then can be used to set the control limits and the charts can be used prospectively to detect process changes
(Phase II). In most practical applications, however, true parameter values are rarely known. Even so, parameters often are estimated, creating random errors affecting run behaviors that diminish chart performance.

The direct solution to this problem is to increase the Phase I sample size to reduce estimate variability. In many cases, however, it may not be possible to wait for a sufficient number of observations to accumulate. This solution would not be viable in short run and startup process settings.

Another remedy involves using a self-starting (SS) method that updates the parameter estimates as new observations are recorded and simultaneously monitors the process to see whether it is in control or not. This method is successful because it produces a stream of standard normal readings from the unknown parameter process data by removing the problem of unknown parameters (Capizzi & Masarotto, 2010; Maboudou-Tchao & Hawkins, 2011; J. H Sullivan & Jones, 2002). The self-starting method creates benefits by reducing the data gathering exercise for Phase I and incorporating a better charting process.

The other alternative to this problem is multivariate change-point (MCP) charting method which is the GLR approach for unknown parameter SQCs. In this case, the parameters \((\mu_1, \Sigma_1)\) for the process after a change are estimated using the maximum likelihood ratio, which undergoes a GLR test for the presence of a change-point against the null hypothesis of no distinction in the sequence of data (Zamba & Hawkins, 2006, 2009). We can test for the presence of a presumed change-point with another generalized likelihood ratio test. This test compares the left and right sections of a sequence, maximized across all possible change-points. The main
statistics are then compared to a sequence of limits, \( h(n) \), which is obtained via simulation, while maintaining the conditional probability of a false alarm (\( \alpha \)) from any observation.

### 4.3 Method

#### 4.3.1 Control Chart Construction

We have chosen to concentrate on the MSS-EWLR chart and MCP chart, techniques that reflect more realistic situations in which the in-control mean vector and covariance matrix are not known \textit{a priori}, yet are able to monitor process parameters simultaneously. In what follows, we describe the statistical construction of these charts. The suggested control charts consider only upper control limits for monitoring a multivariate process. In these multivariate charting procedures, we monitor the significance of the shift magnitude for the mean vector and/or covariance matrix from target or nominal values, and thus the direction of the shift does not play an important role.

We want to monitor a sequence of successive process readings for a \( p \)-dimensional normal distribution where \( X \sim N_p(\mu, \Sigma) \). Let \( x_{i,j}, (i = 1,2,\ldots,n \text{ and } j = 1,2,\ldots,p) \) be \( i^{th} \) observation representing \( j^{th} \) quality characteristic from the process of interest. The distribution shift model is as follows:

\[
X_i \sim \begin{cases} 
N_p(\mu_0, \Sigma_0) & \text{if } i < \tau \\
N_p(\mu_1, \Sigma_1) & \text{if } i \geq \tau,
\end{cases}
\]

where \( \tau \) is the change point between the two segments of data. Both segments are independent and normally distributed. \( \mu_0 \) and \( \Sigma_0 \) are the in-control true mean vector and covariance matrix.
and $\mu_1$ and $\Sigma_1$ are the out-of-control mean (vector) and covariance (matrix) when the process mean shifts.

Under the assumption of no change in a given observation $X$, consider the following hypothesis test

$$H_0: \mu = \mu_0 \text{ and } \Sigma = \Sigma_0.$$ 

The main statistic of the charting process in the MCP chart and MSS-EWLR chart can be calculated with the GLR statistic, which is determined for each chart. The main statistics are compared to a sequence of control limits defined with statistical parameters (mean and variability). If all process values fall under the upper control limits and exhibit no unusual patterns, the variation is considered to be common cause (noise) which is regarded as in-control process. Otherwise, this provides evidence of a special cause of variation (OOC signal) which is likely to be a real change in the health data.

4.3.2 Multivariate Self-Starting Exponentially Weighted Likelihood Ratio (MSS-EWLR) Chart

The MSS-EWLR chart is the integration of the self-starting control chart (Hawkins & Maboudou-Tchao, 2007) with the likelihood ratio test (LRT)-based multivariate exponentially weighted moving average chart (Zhang, Li, & Wang, 2010). To calculate the main statistics of the MSS-EWLR chart, the two steps of self-starting procedure is firstly applied in the dataset to transform the incoming streams of multivariate normal process readings $X \sim N_p(\mu_0, \Sigma_0)$ with unknown parameters into a sequence of mutually independent $R = N_p(0, \sigma_i^2)$ with unknown variances, and then, into a series of multivariate standard normally-distributed data $U \sim N_p(0, I)$. 
The first step is to produce the one-step ahead predicted recursive residuals $R = N_p(0, \sigma^2)$ following mutually independent normal quantities with zero-mean vectors and a diagonal covariance matrix with unknown variances. This provides the sequence

$$r_{i,j} = x_{i,j} - \hat{x}_{i,j},$$

where $\hat{x}_{i,j}$ are the regression-based prediction. This is applied to all observations $i$, but is restricted to $j$ variables as $j = 0, 1, ..., \min(p, n - l)$.

The second step is to produce the standardized recursive residuals $t_{i,j}$ which can be defined based on the first transformation $r_{i,j}$ as

$$t_{i,j} = \frac{r_{i,j}}{\sqrt{\sum_{k=j+1}^{-(i-1)} r_{k,j}}},$$

The standardized recursive residuals $t_{i,j}$ are transformed into standardized normal distributed vectors $u_{i,j}$ as the final self-starting transformation outcomes, given by

$$u_{i,j} = \Phi^{-1}\left[F_{i-j-1}(t_{i,j})\right],$$

where $\Phi^{-1}$ represents the inverse normal, and $F_{i-j-1}(t)$ is the cumulative distribution of $t$ with $i - j - 1$.

The suggested two-step transformation changes original streams of $X$ of vectors with an unknown mean vector and covariance matrix into $p$-variable standardized normal distributed vectors $U$. The transformed vectors ($U$) can be monitored using any desired multivariate monitoring scheme.

Here, for the MSS-EWLR, the multivariate EWMA (MEWMA) is proposed for monitoring mean and variability. Two MEWMA statistics based on the transformed vectors $u_i$ for every observation $i$ are given by
\[ w_i = \lambda u_i + (1 - \lambda) w_{i-1}, \]
\[ v_i = \lambda S_i^* + (1 - \lambda) v_{i-1}, \]

where, \( S_i^* = (u_i - w_i)'(u_i - w_i) \), \( w_0 = 0, v_0 = I_p \), and \( \lambda \) is a smoothing parameter satisfying \( 0 < \lambda < 1 \). In general, a smaller \( \lambda \) leads to quicker detection of smaller shifts.

Then, a GLR statistic with MEWMA statistics \( w_i \) and \( v_i \) is then obtained as follows:

\[ SS-MELR_i = tr(v_i) - log|v_i| + p||w_i||^2. \]

If \( MSS-EWLR > h \), then the process is considered to be out-of-control, where \( h > 0 \) is a threshold to achieve a specified IC ARL.

### 4.3.3 Multivariate Change-Point (MCP) Chart

The main idea behind the multivariate change-point chart (Zamba & Hawkins, 2009) is finding the change-point in the mean and/or covariance (matrix) by sequentially computing the GLR to compare the parameters of the left (\( i \leq \tau \)) and right (\( i > \tau \)) segments.

The likelihood ratio for testing the difference between pre-shift and post-shift data at an assumed change point \( k \) is obtained as follows:

\[ G_{k,n} = \frac{p\log3 \cdot ((n - 1)\log|S_{0,n}| - (k-1)\log|S_{0,k}| - (n-k-1)\log|S_{k,n}|)}{p\log2 + (n - 1)\log(n - 1) - (n-k-1)\log(n-k) - (k-1)\log(k-1) + g_{k,n}}, \]

where

\[ S_{ij} = \sum_{i+1}^j \left( \frac{X_k^\psi}{(j-i)} \right)^{\beta} \left( \frac{X_k^\beta}{(j-i)} \right)^{\psi'}, \]

\[ g_{k,n} = \sum_{j=1}^k (n - 1)\psi \left( \frac{n-j}{2} \right) - (k - 1)\psi \left( \frac{k-1}{2} \right) - (n - k - 1)\psi \left( \frac{n-k-j}{2} \right). \]
when \( \psi(x) \) is the digamma function. The likelihood ratio test statistics \( G_{k,n} \) are maximized across all possible split point \( k \) values, thereby obtaining the generalized likelihood ratio test statistics:

\[
G_{\text{max},n} = \max_{p < k < n - p} |G_{k,n}|.
\]

For each new observation \( n \), \( G_{k,n} \) is calculated for each assumed change point \( k \) which can be range \( p + 1 \leq k \leq (n - p - 1) \) and the \( G_{\text{max},n} \) is determined as the maximum of these \( G_{k,n} \). Then, \( G_{\text{max},n} \) is compared to the control limit (\( h \)). If the change-point chart has OOC signals responding to the changes in the process, the time of occurrence (\( \tau \)) of the shift can be estimated with the index \( k \) corresponding to the maximum likelihood estimate of the change.

### 4.4 Data: Blood Pressure Monitoring Case Study

Biological characteristics can be measured over time in order to assess risk. For example, blood pressure (BP) monitoring is used to assess risk for hypertension as well as stroke, cardiovascular disease and heart related disease, all of which have a high probability of causing early death. It is important to diagnose hypertension and its related complications early in order to take effective medical action; monitoring changes in physiological measurements over time supports these efforts.

Recently, medical researchers have suggested that mean arterial pressure (MAP) as well as systolic blood pressure (SBP) and diastolic blood pressure (DBP) are important barometers for classifying hypertension and its related complications (Benetos, Laurent, Asmar, & Lacolley, 1997; Blacher, et al., 2000; Safar, 1989). In addition, it has been recognized that BP levels are
highly correlated (Chobanian, et al., 2003; Darne, Girerd, Safar, Cambien, & Guize, 1989; O'Rourke, 1982); thus, monitoring BP levels separately without accounting for the correlation among them would be discordant with process control realities.

Figure 4.1. Ambulatory monitoring outcomes for SBP, DBP and MAP.

Based on the importance of monitoring BP levels over time, we applied the suggested charting methodologies (MCP and MSS-EWLR), which explain changes in both location and variability. We used the ambulatory BP data set from Hawkins and Maboudou-Tchao (2008) containing information for three variables: SBP, DBP, MAP. For more consistent results, we generated 50 observations based on the in-control parameters of multivariate normal distributed process from the long historical sequences of the observed person using the mean vector

$$\mu_0 = [126.61 \quad 77.48 \quad 97.997]'$$,
and covariance matrix

\[
\Sigma_0 = \begin{bmatrix}
15.04 & 8.66 & 12.04 \\
8.66 & 5.83 & 7.50 \\
12.04 & 7.50 & 10.57
\end{bmatrix}
\]

Next, 24 measurements based on ambulatory monitoring are added so that weekly measured ambulatory BP levels were observed as given in Figure 4.1.

4.5 Results

4.5.1 Case Study

We applied MCP and MSS-MERL charts to these series to identify when the BP characteristics changed, thus increasing or decreasing health risk. In both charts, we set the smoothing constant \( \lambda \) to 0.1 and the IC ARL to 500, which implies that the false alarm level was 0.02. The control limits were \( h = 1.246 \) for the MSS-EWLR and \( h = 6.676 \) for the MCP.

MSS-EWLR chart requires going through two transformation steps before applying EWLR statistics. Table 4.1 lists the first few cases, showing the original data, the recursive residuals \( R \) as the first transformation outcomes, and then, a sequence of mutually independent \( p \) vectors \( U \) as the second transformation. Recursive residuals \( \{R\} \) can be defined from the \((p+1)\)th observations because they are obtained from the one-step ahead predication with previous variables. It completes the objective of the first transformation step which yields independent normal zeros mean quantities.
Table 4.1. First few cases of the BP Profile data showing the self-starting procedures and MSS-EWLR statistics

<table>
<thead>
<tr>
<th>Observation</th>
<th>SBP</th>
<th>DBP</th>
<th>MAP</th>
<th>MSS_EWLR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>122.95</td>
<td>75.64</td>
<td>96.98</td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>124.01</td>
<td>77.08</td>
<td>96.28</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>6.36</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>128.85</td>
<td>79.77</td>
<td>101.14</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>5.33</td>
<td>-0.81</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>118.52</td>
<td>72.49</td>
<td>91.88</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>-1.33</td>
<td>1.84</td>
<td>-0.78</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>-0.23</td>
<td>2.28</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>-0.20</td>
<td>1.12</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>127.53</td>
<td>77.89</td>
<td>99.80</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>1.72</td>
<td>0.17</td>
<td>-0.17</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.35</td>
<td>0.12</td>
<td>-0.22</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>0.32</td>
<td>0.11</td>
<td>-0.17</td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>123.59</td>
<td>76.42</td>
<td>96.40</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>2.64</td>
<td>-0.45</td>
<td>-1.04</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.62</td>
<td>-0.38</td>
<td>-1.85</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>0.57</td>
<td>-0.35</td>
<td>-1.27</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>130.87</td>
<td>80.05</td>
<td>99.63</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>-1.02</td>
<td>0.31</td>
<td>-0.70</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>-0.26</td>
<td>0.30</td>
<td>-0.93</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>-0.24</td>
<td>0.28</td>
<td>-0.80</td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>123.29</td>
<td>76.61</td>
<td>95.38</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>1.29</td>
<td>-0.28</td>
<td>-0.55</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.35</td>
<td>-0.30</td>
<td>-0.73</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>0.33</td>
<td>-0.28</td>
<td>-0.67</td>
</tr>
<tr>
<td>9</td>
<td>X</td>
<td>126.64</td>
<td>76.87</td>
<td>97.95</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>2.70</td>
<td>0.69</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>0.78</td>
<td>0.81</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>0.74</td>
<td>0.75</td>
<td>0.44</td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td>122.97</td>
<td>75.82</td>
<td>95.54</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>-2.55</td>
<td>-0.41</td>
<td>0.32</td>
</tr>
<tr>
<td></td>
<td>T</td>
<td>-0.76</td>
<td>-0.49</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>U</td>
<td>-0.72</td>
<td>-0.47</td>
<td>0.46</td>
</tr>
</tbody>
</table>
In order to get a stream of standardized normal distribution quantities \( (U) \), the studentized recursive residuals \( (T) \) are evaluated. The final self-starting transformation outcomes \( (U) \) are converted into sequence of standard normal by applying cumulative distribution function of the student t distribution and inverse normal function. The sequence of observations \( (U) \), following independent standard normal distribution, can be monitored using EWLR chart. Two MEWMA statistics are evaluated with smoothing parameter \( (\lambda) \) as 0.1 and final the MSS-EWLR charting statistic is obtained.

Applying the MCP charting procedure with BP profiles (SBP, DBP, and MAP), the main statistics \( (G_{\text{max}}) \) are evaluated with the likelihood ratio \( (G_{k,n}) \) with the sequence of length \( n \) and an assumed change point \( k \). The main statistics \( (G_{\text{max},n}) \) of MCP methodology can be obtained by maximizing the likelihood ratio \( (G_{k,n}) \) for possible change-points \( (k) \). All the values of \( G_{\text{max},n} \) are compared to control limit \( h \) in order to get the signal of change as well as its pertinent change-point yielding maximum of the likelihood \( (G_{k,n}) \) as seen in Table 4.2.

Figure 4.2 shows the result of applying the MSS-EWLR and MCP charts to BP profiles data. Examining the outcomes of two suggested methodologies, we can see that an OOC behavior was signaled simultaneously in both charts at the 58th observation. Based on the performance of SS-MERL chart in simulation, we expected it to provide an earlier OOC signal based on changes in ambulatory monitoring. In these charts, OOC signals were continuously observed after the 58th observation and we can say that the patient experienced some major changes regarded as special causes at that time.
Table 4.2. The $G_{\text{max}}$ statistics of the MCP chart and its corresponding change point (CP) of the BP profiles.

(The rows in bold indicate where the $G_{\text{max}}$ exceeded control limit, $h = 6.676$.)

<table>
<thead>
<tr>
<th>Observation</th>
<th>SBP</th>
<th>DBP</th>
<th>MAP</th>
<th>$G_{\text{max}}$</th>
<th>CP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>122.95</td>
<td>75.64</td>
<td>96.98</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>124.01</td>
<td>77.08</td>
<td>96.28</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>128.85</td>
<td>79.77</td>
<td>101.14</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>118.52</td>
<td>72.49</td>
<td>91.88</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>127.53</td>
<td>77.89</td>
<td>99.80</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>123.59</td>
<td>76.42</td>
<td>96.40</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>130.87</td>
<td>80.05</td>
<td>99.63</td>
<td>0.00</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>123.29</td>
<td>76.61</td>
<td>95.38</td>
<td>4.99</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>126.64</td>
<td>78.87</td>
<td>97.95</td>
<td>3.78</td>
<td>5</td>
</tr>
<tr>
<td>10</td>
<td>122.97</td>
<td>75.82</td>
<td>95.54</td>
<td>5.95</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>55</td>
<td>133.07</td>
<td>81.29</td>
<td>103.53</td>
<td>3.31</td>
<td>51</td>
</tr>
<tr>
<td>56</td>
<td>127.67</td>
<td>79.63</td>
<td>100.10</td>
<td>2.07</td>
<td>52</td>
</tr>
<tr>
<td>57</td>
<td>130.43</td>
<td>81.06</td>
<td>102.41</td>
<td>3.62</td>
<td>53</td>
</tr>
<tr>
<td>58</td>
<td>124.87</td>
<td>77.68</td>
<td>87.46</td>
<td>10.79</td>
<td>54</td>
</tr>
<tr>
<td>59</td>
<td>129.39</td>
<td>78.73</td>
<td>99.31</td>
<td>10.32</td>
<td>55</td>
</tr>
<tr>
<td>60</td>
<td>127.81</td>
<td>78.73</td>
<td>98.96</td>
<td>10.41</td>
<td>56</td>
</tr>
<tr>
<td>61</td>
<td>129.98</td>
<td>78.27</td>
<td>100.03</td>
<td>11.99</td>
<td>57</td>
</tr>
<tr>
<td>62</td>
<td>128.30</td>
<td>76.63</td>
<td>98.92</td>
<td>9.86</td>
<td>57</td>
</tr>
<tr>
<td>63</td>
<td>124.36</td>
<td>75.25</td>
<td>96.77</td>
<td>8.02</td>
<td>57</td>
</tr>
<tr>
<td>64</td>
<td>131.20</td>
<td>78.15</td>
<td>100.99</td>
<td>7.55</td>
<td>57</td>
</tr>
<tr>
<td>65</td>
<td>127.08</td>
<td>76.43</td>
<td>97.17</td>
<td>7.09</td>
<td>57</td>
</tr>
<tr>
<td>66</td>
<td>123.84</td>
<td>75.74</td>
<td>94.55</td>
<td>6.43</td>
<td>57</td>
</tr>
<tr>
<td>67</td>
<td>125.49</td>
<td>75.99</td>
<td>96.13</td>
<td>6.08</td>
<td>57</td>
</tr>
<tr>
<td>68</td>
<td>121.08</td>
<td>73.58</td>
<td>92.32</td>
<td>5.57</td>
<td>57</td>
</tr>
<tr>
<td>69</td>
<td>121.46</td>
<td>73.95</td>
<td>91.39</td>
<td>5.41</td>
<td>57</td>
</tr>
<tr>
<td>70</td>
<td>128.09</td>
<td>79.59</td>
<td>98.55</td>
<td>4.67</td>
<td>57</td>
</tr>
<tr>
<td>71</td>
<td>127.76</td>
<td>77.35</td>
<td>98.46</td>
<td>4.48</td>
<td>57</td>
</tr>
<tr>
<td>72</td>
<td>125.20</td>
<td>75.73</td>
<td>96.09</td>
<td>4.36</td>
<td>57</td>
</tr>
<tr>
<td>73</td>
<td>128.21</td>
<td>77.83</td>
<td>101.14</td>
<td>4.23</td>
<td>57</td>
</tr>
<tr>
<td>74</td>
<td>125.80</td>
<td>76.39</td>
<td>95.09</td>
<td>4.22</td>
<td>57</td>
</tr>
</tbody>
</table>
The change-point (τ) information in the MCP method confirms BP process changes after the 58th observation, because the change-point (τ) locations indicating the onset of change were observed at the 57th observation after the 65th observation.

![Figure 4.2. MCP (left) and MSS-EWLR (right) for ambulatory monitoring outcomes.](image)

Examining the outcomes of two suggested methodologies, we can see that the OOC behavior was signaled simultaneously in both charts at the 58th observations. The diagnostic analysis of the OOC signals with multivariate change point (MCP) and multivariate self-starting exponentially weighted likelihood ratio (MSS-EWLR) charts is an important consideration. The main idea of the MCP chart is to estimate the change point (τ), a major split between in-control and out-of-control process by sequentially computing the GLR statistics for the left and right segments for potential change-point (τ). The location of change-point maximizing the GLR test for potential is found in the 57th observation. The information of change point (τ) enables estimation of the parameter information about pre-shift and post-shift data based on change point (τ) information and to summarize the MCP chart model as follows:
\[ X_i = \begin{cases} N_3(\mu_0, \Sigma_0) & \text{if } i < 57 \\ N_3(\mu_1, \Sigma_1) & \text{if } i \geq 57 \end{cases} \]

where, mean vectors of observation \( X_i \) are

\[ X_{\mu_0} = [127.40 \quad 78.30 \quad 98.83]' \], and \( X_{\mu_1} = [126.46 \quad 75.82 \quad 96.66]' \),

covariance matrices of observation \( X_i \) are

\[ X_{\Sigma_0} = \begin{bmatrix} 15.41 & 8.94 & 12.34 \\ 8.94 & 5.99 & 7.56 \\ 12.34 & 7.45 & 10.74 \end{bmatrix} \], and \( X_{\Sigma_1} = \begin{bmatrix} 7.945 & 4.06 & 8.30 \\ 4.06 & 2.84 & 3.46 \\ 8.30 & 3.46 & 13.34 \end{bmatrix} \),

and, correlation matrices of observation \( X_i \) are

\[ X_{\text{corr}_0} = \begin{bmatrix} 1 & 0.930 & 0.959 \\ 0.930 & 1 & 0.941 \\ 0.959 & 0.941 & 1 \end{bmatrix} \], and \( X_{\text{corr}_1} = \begin{bmatrix} 1 & 0.858 & 0.828 \\ 0.858 & 1 & 0.633 \\ 0.828 & 0.633 & 1 \end{bmatrix} \).

The parameter summaries between pre-shirt and post-shift proves that the signal may have resulted from a shift in the diastolic blood pressure and mean arterial pressure and a distortion in the correlation structure between diastolic and mean arterial pressure. The diagnosis results for OOC signals in MCP chart can be also found in the MSS-EWLR chart.

Here, the observations used for the main-statistics are a series of multivariate standard normally distributed data \( U \) transformed from incoming streams of multivariate process readings \( X \) with unknown parameters. Noting that vectors \( U \) are multi-standardized transform, the mean vector of \( U \) should be composed of zeros and the covariance matrix of \( U \) should be the identity. The comparison between the actual statistics of observations \( U \) and its expected values can provide evidence for the shift and its directionality. The parameter information of \( U \) for the post-shift data (\( i \geq 57 \)) can be summarized as follows:

\[ U_{\mu_1} = [-0.108 \quad -0.587 \quad -0.653]' \],

76
Comparing the post-shift parameter estimates obtained responding to OOC signals and the expected parameter information from zero mean vector and identity covariance matrix, the shifts in mean are observed in diastolic blood pressure and mean arterial pressure. Also, there are changes in correlations jointly combined with diastolic and/or mean-arterial pressure.

Statistically, we have identified changes in blood pressures and the possible direction and factors of the changes. In the medical field, statistical significance does not necessarily translate into clinical importance. However, the statistical information can provide evidences to look into the health conditions of the observed patient especially in terms of diastolic blood pressure and mean arterial pressure, noting the importance of monitoring blood pressure levels which can predict to progression of the heart related disease as well as hypertension and its complications.

\[ U_{\Sigma_1} = \begin{bmatrix} 0.593 & -0.130 & 0.474 \\ -0.130 & 1.171 & -1.042 \\ 0.474 & -1.042 & 4.924 \end{bmatrix} \]

4.5.2 Simulation Study: Comparisons of MSS-EWLR and MCP

The performance of the proposed charts is assessed by ARL of the chart at different levels of shifts in both mean vector and/or covariance matrix. ARL is the average number of points plotted on a control chart before a point signals an out-of-control (OOC) condition. When a special cause does occur, the ARL suggests how quickly the chart responds to process changes. The shorter the ARL for out-of-control conditions, the better the chart’s performance.

We assessed the performance of the proposed MSS-EWLR chart and the MCP chart for individual observations using an in-control ARL of 500 runs for different \( \tau \) levels. The control limits corresponded to a Type I error probability of 0.002. To assess the impact of changes in the
mean vector alone, in the covariance matrix alone, and in both the mean vector and covariance matrix, we calculated up to 1,000 ARL simulations, each 10,000 sequences in length using dimension $p = 2$. Each simulation was studied by generating a sequence of a standard normal data. Random samples from fixed multivariate normal distributions were generated for various change-points ($\tau$) as 20, 50, 100, and 200 at which parameter changes (in the mean vector and/or covariance) occurred.

Figure 4.3. The effect of mean shift: ARL OOC performance study of the MSS-EWLR (Left) and MCP (Right) with different $\tau$ (20, 50, 100, 200) in the mean shift change: $\delta = (0.5, 1, 2, 4)$. Each line represents a specific mean shift, starting with 0.5 at the top and 4 at the bottom.
Figure 4.4. The effect of shift on covariance shift: ARL OOC performance study of the MSS-EWLR (Left) and MCP (Right) with different $\tau$ (20, 50, 100, 200) in covariance shift changes: $\rho = (0.2, 0.4, 0.6, 0.8)$. Each line represents a specific covariance shift, starting with 0.2 at the bottom and 0.8 at the top.

Figure 4.5. The effect of a shift in mean vector coupled with a shift in covariance matrix: ARL OOC performance study of the MSS-EWLR (Left) and MCP (Right) with different $\tau$ (20, 50, 100, 200) in mean and covariance shifts: $(\delta, \rho) = [(0.5, 0.6), (0.5, 0.2), (1, 0.6), (1, 0.2)]$. Each line represents a specific coupled mean and covariance shift, starting with (0.5, 0.6) at the top and (1, 0.2) at the bottom.

In our simulation, we considered the performance changes based on mean vector and/or covariance matrix shifts and impact of change-points ($\tau$). For the mean shift ($\delta$), we considered shifting a single component in the mean vector. For the covariance matrix, we multiplied by $\rho$. Then, we considered various combinations of $\delta$ and $\rho$: $\delta = (0.5, 1, 2, 4)$, $\rho = (0.2, 0.4, 0.6, 0.8)$. After these changes, we created charts to identify change-point location signals ($\tau = 20, 50, 100, 200$).

The overall patterns of outcomes in Figures 4.3, 4.4, and 4.5 show that ARLs decreased and approached to ones for the large shifts, while the process mean vector and/or covariance matrix shifted. Both charting procedures responded more effectively as the number of change-points ($\tau$)
increased, while the ARL decreased as the value of $\tau$ increased. Both procedures included the impact of the in-control process run time before the shift. The suggested methodologies estimated the in-control parameters for every new monitoring process, and long in-control runs better estimated parameters leading to rapid response for OOC signals.

**4.5.3 Discussion**

The main objective in health monitoring is to identify abnormal trends in physiological variables. Most previous monitoring guidelines tend to ignore the variability inherent in physiological variables as well as errors involved in making clinical measurements. Since all medical conditions or outcomes can be measured either directly or via biomarkers, it is expected that monitoring procedures will result in errors or uncertainties among measurements over time, even if the true underlying condition is stable. To overcome such limitations, SQC tools are becoming widely accepted in the healthcare domain as a means of monitoring processes and outcomes by finding evidence for special causes of variation, while also identifying common causes in the process.

However, SQC tools still fall short in practical healthcare monitoring applications. First, more than one quality characteristic must be tracked in clinical settings; such characteristics are usually correlated so both variability and mean levels must be monitored. Generally, patient health conditions are diagnosed based on mean levels. However, increased variability can predict illness, even if the mean level is normal.

Second, it is difficult to gather sufficient in-control process data regarded in clinical settings. Even if the datasets could be prepared to observe the data patterns, we could not assume the
collected data to be in-control. However, when such datasets are used, the necessary in-control parameters defining the control limits are not estimated.

In light of these challenges faced in healthcare monitoring, we introduced two multivariate SQC charts (MCP and MSS-EWLR) that do not require in-control parameter information such as mean (vector) and covariance matrix to be specified, yet maintain the ability to simultaneously monitor the mean and process variability for multivariate processes using one single chart. Additionally, these two proposed charts can be easily designed and constructed compared to other SQC methodologies, because there is no need to gather in-control process data. Also, the suggested charts in the chapter provide quite robust and satisfactory performance in various cases, because it inherits the powerful properties of the GLR test.

It is important to have earlier detection ability regardless of shift size or location (mean and/or covariance matrix) because monitoring in a clinical setting requires detecting subtle to moderate changes in the true underlying condition of a patient. In line with practical needs, we used simulations to compare chart performance. Previous studies have proven that GLR-based MEWLR and MCP charts are most optimal for a variety of shifts, compared to the MCUSUM, MEWMA or MEWMC which are known to be useful for assessing small shifts quickly. Based on simulation results, the MSS-EWLR chart performed better for small to moderate shifts in the mean vector and/or covariance matrix.

One difficulty encountered in multivariate settings is practical interpretation of OOC signals. After the OOC signal, a complex diagnosis is required to see what characteristics of the process have gone out of control. The standard approach is to plot univariate charts, which can reduce the number of false alarms associated with using simultaneous univariate control charts. Another
way to diagnose an OOC signal is to decompose the main statistic into $p$ components that reflect the contribution of each individual variable. They are indicators of the relative contributions of each variable to the overall statistics, and require quite extensive computations.

Unlike the other general multivariate SQC tools, including the MSS-EWLR chart, MCP method includes a diagnostic tool. Whenever MCP calculates main statistics of new observations, it involves change-point estimates for major splits between pre-shift and post-shift data. The change-point ($\tau$) information enables parameter information for the observations before $\tau$ and after $\tau$ to be compared, supporting the charting diagnosis process to identify factors responsible for the OOC signals.

SQC tools fundamentally rely on the assumption that the observations must be independently and normally distributed. Using SQC tools to monitor data that violate these general assumptions presents some problems. It interferes not only with tracking movements, but also in obtaining a timely and sensitive signal for the process. To deal with the auto-correlation often seen in health data, a vector autoregressive moving average (VARMA) model can be used. The residuals between original data and VARMA model forecasts are used in SQC tools instead of original data. In addition, normality transformation (taking log or inverse) can be considered before using SQC directly when the normality assumption is violated.

It is important to consider the number of variables in multivariate monitoring. Generally, multivariate monitoring charts are used to avoid excessive false signals associated with using separate univariate charts. However, it is also well known that a charting procedure can perform well as long as the number of process variables ($p$) to be monitored is not large. ARL
performance for process changes increases as \( p \) increases, because the shift cannot be exposed well.

Also, the suggested SQC charts in this chapter involve the minimum number of observations to calculate the main statistics. MSS-EWLR requires a minimum sample \( p \) to calculate the one-step ahead recursive prediction of self-starting methods, while MCP requires \( 2p+1 \), which is the minimum sample size required for both divided covariance matrices to have a chance of being singular. It means that the suggested chart can start charting a process after a minimum number of observations. It also implies that before charting a process, the minimum sample increases as the number of variables to be monitored increases. When the process variables are large, a clinician (or chart designer) can consider using either principal component analysis for sub-dimensions, or a MSS-EWLR chart, which requires fewer observations before start-up while providing better monitoring capability.

Regular monitoring of health outcomes has been recommended in recent national and international guidelines for patient care. However, the collected data are rarely shared with patients to help them understand their conditions, or used by clinicians to modify treatment plans. SQC tools provide visual data displays; control charts not only identify trends, but also clarify when patients should seek additional clinical advice. While rapid major change patterns can be observed easily even without these sophisticated tools, slow and steady changes often cannot be detected easily. SQC tools can be tremendously useful in healthcare monitoring because even subtle change patterns can be detected.

Control charts allow us to focus on specific changes in health marker values that indicate new disease patterns, while ignoring random variation. Using SQC applications in health
monitoring can improve both organizational performance and patient outcomes. These charting processes can be used to monitor quality of care and health services, and to educate people how to perform ambulatory monitoring or read charts.
While interest in individual patient monitoring is growing, quantitative research using national health surveys is often limited to analysis of population groups. In response, we propose a framework for transforming cross-sectional data into repeated measurement data that describes individual patient profiles. With modest investments in data preparation and statistical analysis, profile data estimation (PDE) can create repeated measurements of health conditions estimates along with regression outcomes. We applied this methodology to national health and nutrition examination survey (NHANES) data on hypertensive patients in order to create synthetic patient and cohort blood pressure profiles for non-diabetic normal-weight, over-weight, and obese people. The proposed PDE transformation process can be used to perform a pilot validation of new trials through extraction of cross-sectional data which may lead reduced time and data collection costs, and better comprehension of patient health statuses, which ultimately supports a higher quality of care.

5.1 Introduction

Hypertension is a national health concern in the United States. More than one in three Americans suffer from hypertension, and the disease is associated with about one in six deaths (Fields, et al., 2004; Institute of Medicine, 2010). Elevated blood pressure (BP) level is a primary risk factor for
cardiovascular disease (CVD), heart attack, heart failure and stroke, all of which have high probability of causing of death (Chobanian, et al., 2003; Stamler, Stamler, & Neaton, 1993; Stamler, Vaccaro, Neaton, & Wentworth, 1993; Wang, Staessen, Franklin, Fagard, & Gueyffier, 2005; Wang, et al., 2006). Conversely, small reduction in BP levels result in a large reduction in the risk of hypertension and its comorbidities which are the cause of the early death (Terry, et al., 2005; Vasan, et al., 2001). In this regard, there is tremendous need for monitoring BP levels for better improvement of hypertension management.

Considering the chronic disease properties of long-term latency periods and relapsing conditions, it is essential to monitor clinical variables sequentially and periodically to make more accurate health diagnosis and better chronic disease management. Hypertension, or elevated BP, is one of the most common chronic diseases. An early diagnosis would provide opportunities for proactive steps. Therefore, it is important to observe patterns of the BP in order to assess risk for hypertension and its complications; monitoring changes in BP measurements over time supports these efforts.

However, repeated measurements of health outcome information with regard to chronic disease generally costs money and time in proportion to the frequency and period of data collection. To mitigate these costs, researchers may seek to outline possible scenarios of the research and try a simulated study with an alternative dataset for validating the new trials. Most accessible national surveys are based on the cross-sectional data that depicts a snapshot of the population health conditions and are designed to make county-level estimates of health status and health service utilization. Even if national surveys are performed periodically with the same individuals they do not provide linking identifiers in order to protect their privacy and
confidentiality. This limits the ability to analyze the dynamic changes of individuals over time with national data. In response to ongoing need for repeated measurement data of health outcomes, we suggest a profile data estimation (PDE) methodology to estimate the profiles of health outcomes of chronic disease including hypertension. The PDE methodology is designed to create such profile estimates of health outcomes containing temporal properties which enables the assessment and refinement of potential studies prior to implementation.

There are many risk factors that impact BP levels: obesity, exercise, genetics, medical history, smoking and alcohol consumption (Berger, et al., 1999; Bowman, Gaziano, Buring, & Sesso, 2007; Burke, et al., 2008; Choi, et al., 2005; Juvela, Hillbom, Numminen, & Koskinen, 1993; Lavie, et al., 2000; Leif, 2009; Rimm, Williams, Fosher, Criqui, & Stampfer, 1999). Among them, obesity is one of the most important modifiable risk factors for hypertension among populations as demonstrated in large number of cross-sectional and prospective epidemiological studies (Burke, et al., 2008; Choi, et al., 2005). These studies suggest that excess body weight is positively associated with BP levels in both men and women, irrespective of age.

This chapter describes the PDE method for producing BP profile estimates. We create estimated repeated measurements of blood pressure in normal-weight, over-weight and obese subjects. The degree of obesity is defined as BMI, calculated with height and weight measurements. We defined obese person as ever having at least 30 BMI and over-weight person as more than 25 BMI. We can examine the effects of excess body weight on BP by comparing the average BP levels with two cohort groups. We can anticipate that BP levels of obese persons have higher mean values in terms of SBP and DBP rather than normal-weight or over-weight
persons. The comparison of mean BP levels can suggest the validation and utility of the PDE methodology in the healthcare domain.

The remainder of the chapter is organized as follows. In the next section, we provide a review of the relevant literature. Then we describe the NHANES data sources. Following this, we outline the technical details of the PDE methodology in order to create repeated measurements estimates of BP profiles among normal-weight, over-weight and obese persons. We illustrate the PDE methodology using BP profiles with these three cohorts. In the final section, we discuss and make several suggestions with regards to the potential extensions with PDE methodology.

5.2 Literature Review

Repeated measures data provide information on a variable or a set of variables repeatedly under different experimental conditions. In biological and biomedical applications, it is common to observe repeated measurements of bio-markers taken on number of individuals. Compared to the cross-sectional or descriptive studies in which the researcher observes only one value per person, repeated measurement studies have multiple observations per person. In the biomedical area repeated measures studies are beneficial for explaining the interventions or experiments. For example, measurements are taken on each of a number of biomarkers such as blood pressure or blood samples on a regular time interval bases to estimate the effect of medications (such as antihypertensive medicine) with placebo-controlled cohort studies. Subjects are randomly assigned to receive either treatment or control condition. The serial blood pressure measurements
are made over time for participants and the mean of blood pressure is compared between treatments.

Researchers commonly design repeated measures analysis studies to examine intervention effects. Since chronic diseases can be worsened or improved by dietary and other lifestyle factors, there have been often studies where patients are asked to modify behaviors. For example, patient education and self-management programs may suggest the adoption of exercise regimens, the use of cognitive symptom management, following dietary approach to stop hypertension (DASH) diet plans, managing fatigue and sleep patterns, strategies for medication compliance, and limiting smoking or alcohol consumption (NHLBI, 2006). These can improve hypertension or its related complications while reducing health care costs in populations with diverse chronic diseases (Lorig, et al., 2001; Newman, Steed, & Mulligan, 2004; Nuovo & Levich, 2007; Scanlon, et al., 2000; Warsi, Wang, LaValley, Avorn, & Solomon, 2004; Yamaguchi, et al., 2003). In line with these potential impacts in chronic disease management, many repeated measures analysis studies provide outcome comparisons based on statistical analysis (student \( t \) test for continuous variables, \( chi-square \) test for categorical data) between patients who have completed intervention programs and those who have not, in terms of health status, self-efficacy, and healthcare utilization.

Repeated measures analysis studies provide precise estimates for trends and rates of change over time. As criteria of chronic disease, physiological measurements can be affected by age. Information about general physiological patterns in the overall population or a subgroup of research interest enables researchers to make predictions and understand the nature of chronic disease. Common age-related physiological patterns that are tracked in populations of interest
include blood pressure. In several studies, researchers have used multivariate linear regression models to assess physiological variations within and across patients, and modeled growth curves to examine and contrast the demographic effects on physiological measurements (Franklin, et al., 1997; Ronnback, et al., 2004; Wang, et al., 2006; Yamaguchi, et al., 2003).

Blood pressure is impacted by various multi-disciplinary inputs, including social and psychological factors. Thus, some studies suggested anger and anxiety as the psychosocial main factors that are associated with the likelihood of developing high blood pressure (Friedman, et al., 2001; Gerin, et al., 2006; Schwartz et al., 2003). It is widely believed that situational and psychological factors are important in the determination of blood pressure in clinics. Psychological variables such as anger, anger expression, anxiety, symptoms of psychological distress, locus of control, trait anger score have been investigated through previous research (Cochrane, 1973; Markovitz, Matthews, Kannel, Cobb, & D'Agostino, 1993; Sparrow, Garvey, Rosner, & Thomas, 1982). To illustrate these associations, logistic regression analysis is used to estimate the association of hypertension with each personality/psychological characteristics.

Blood pressure has inherent variability that can impact improper diagnosis. To deal with the variability of blood pressure, researchers have suggested two main repeated measure designs to yield accurate and timely diagnosis. The first approach is to find the frequency and interval of the blood pressure measurements in clinics. Traditionally, the diagnosis associated with hypertension and its related complications is made based on the blood pressure measurements in clinical settings. However, blood pressure involves the common cause variability of fluctuations in a patient’s biological process as well as the assignable cause variability of white-coat effect showing elevated blood pressure in a clinical setting but not in other settings. The variability of
blood pressure in clinical settings may contribute to difficulties in getting a reliable diagnosis of hypertension stages. To deal with the variability of blood pressure in clinical settings, researchers have used the repeated measures design to investigate how many blood pressure measurements are required in diagnosing hypertensive stages. The difference between pairs of subsequent measurements and the average of the serial BP measurements are compared with conceptual average blood pressure in terms of the statistical significances (Bonita, et al., 2008; Brueren, Petri, Van Weel, & Van Ree, 1997; Parati, et al., 2010; Perry & Miller, 1992; Sever, et al., 1993; Struyvenberg, 1990).

The second approach is to find a credible way to measure blood pressure that can explain the progress of the disease as well as the current status of disease. There are three ways that are often considered: clinical blood pressure, ambulatory blood pressure, and self-measurement of blood pressure at home. The issue is which way is most useful for diagnosing and managing hypertension and its related complications. Therefore, researchers have investigated multiple design settings of repeated measurements in terms of the number of measurements, the intervals to measure, time to measure (day or night) (Brueren, et al., 1997; Coats, Radaelli, Clark, Conway, & Sleight, 1992; Dendool, 1989; James, et al., 1988; Mengden, Battig, & Vetter, 1991; Sakuma, et al., 1997; Stergiou, et al., 2002; Stergiou, Skeva, Zoubaki, & Mountokalakis, 1998; Trazzi, et al., 1991). To illustrate the significance and associations with proper diagnosis of hypertensive stages, ANOVA and student t test have been used to compare the average of the blood pressure repeated measurements and the standard deviation of difference between repeated measurements.
5.3 Data

The main data sources used for this study were National Health and Nutrition Examination Surveys (NHANES) for 1999-2008. The NHANES program of the National Center for Health Statistics, Centers for Disease Control and Prevention, includes a series of cross-sectional, nationally representative health examination surveys beginning in 1960. Each cross-sectional survey provides national estimates for the U.S. population at the time of the survey, enabling the examination of trends over time.


Each survey was conducted with a nationally representative sample of the U.S. civilian non-institutionalized population using a complex, stratified, multistage probability cluster sampling design. In the NHANES, African Americans, Mexican Americans, and people aged 60 years or older were oversampled in order to provide better estimates for these groups. To adjust for oversampling and nonresponsive bias, sampling weights were used in the calculation of means and in regression analysis (Cheung, et al., 2006; Nickolas, Frisch, Opotowsky, Arons, & Radhakrishnan, 2004).
5.4 Methods

Aging affects all physiological processes. The rapidity of functional decline varies by organ system, but the aging rate remains constant within a given system (Boss & Seegmiller, 1981; Franklin, et al., 1997; Janssens, Pache, & Nicod, 1999; Lazarus & Harridge, 2010; Weinstein & Anderson, 2010; Woo, Ho, Lau, Chan, & Yuen, 1995). Generally, aging in the entire body can be explained by two factors: normal functional attrition occurring in all people with advancing age, and function loss due to disease over the lifespan. This implies that the rate of aging for the average person is fixed and its rate depends on the age. Considering this evidence, it can be inferred that people can experience the same patterns of physiological change based on advancing age, as well as same rate of aging under analogous health conditions.

The PDE methodology involves evaluating average physiological variables of people with analogous health conditions and calculating the changing rate of aging and patterns among the physiological variables. Averaging the physiological variables by age yields patterns as well as a rate of aging, which can be explained using a regression model. The regression model is then fit to one specific person matching the research interest, and used to find other people whose health conditions seem to be analogous to the selected person using Euclidean norm of residuals.

The estimates produced using this approach are considered “synthetic” because we start with survey data designed for estimating age-associated physiological patterns with regression models at the population level. We use the pattern of age-related associations for the overall population in the statistical regression modeling. Based on the regression outcomes, we can derive estimated BP profiles for individual patients, thus satisfying the research interest.
For example, the PDE methodology produces BP profiles for non-diabetic people who are normal-weight, over-weight, and obese by considering the age-associated patterns of systolic blood pressure (SBP), diastolic blood pressure (DBP), and glyco-hemoglobin (HbA1c) patterns. Obesity classification (normal-weight, over-weight, and obese) are determined using weight and height to calculate the body mass index (BMI), a measure of the amount of body fat. Normal-weight is defined as a BMI of 18.5 to 24.9 kg/m\(^2\), over-weight as a BMI of 25 to 29.9 kg/m\(^2\), and obese as a BMI of at least 30 kg/m\(^2\).

Figure 5.1. Overview of the profile data estimation (PDE) methodology as applied to blood pressure (BP).
Figure 5.1 presents an overview of the four major steps of our PDE methodology: consolidating data, sorting data, regression modeling and creating patient profiles based on health outcome estimates. In this investigation, we explore the PDE methodology by estimating BP profiles for non-diabetic people in their 50s in three cohorts based on the BMI categories of normal-weight, over-weight and obese. The relationship between SBP, DBP, and HbA1c physiological patterns and age among non-diabetic people has been well illustrated in previous studies (Burt, et al., 1995; Franklin, et al., 1997; Franklin, et al., 2009; Pani, et al., 2008). Age-associated physiological patterns are considered in the SBP, DBP, and HbA1c patterns. The final cohorts each contain BP profiles for 30 people whose BP levels were measured 75 instances over time. Considering the effect of excess weight on BP levels, differences in the estimated BP profiles based on BMI categories can be expected.

5.4.1 Creating a Consolidated Dataset

To use the PDE methodology to explore our research interest in BP, we began by creating a consolidated dataset from the NHANES (1999-2008). The continuous NHANES report that begins with the NHANES (1999-2000) is based on data from approximately 4,000 adult men and women. We limited our eligible study population to approximately 2,400 individuals aged ≥ 20 years who were regarded as non-diabetic in each NHANES survey, as shown in Table 5.1. Ten years of data covering five NHANES surveys (1999-2008) were used to create adequate sample sizes in order to better estimate patient and cohort profiles.
Table 5.1. Number of non-diabetic participants by age and BMI classification in NHANES (1999-2008).

<table>
<thead>
<tr>
<th>AGE</th>
<th>NHANES (Non-Diabetic Population)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NORMAL-WEIGHT</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td>190</td>
</tr>
<tr>
<td>25-29</td>
<td>141</td>
</tr>
<tr>
<td>30-34</td>
<td>123</td>
</tr>
<tr>
<td>35-39</td>
<td>122</td>
</tr>
<tr>
<td>40-44</td>
<td>102</td>
</tr>
<tr>
<td>45-49</td>
<td>96</td>
</tr>
<tr>
<td>50-54</td>
<td>79</td>
</tr>
<tr>
<td>55-59</td>
<td>60</td>
</tr>
<tr>
<td>60-64</td>
<td>79</td>
</tr>
<tr>
<td>65-69</td>
<td>63</td>
</tr>
<tr>
<td>70-74</td>
<td>83</td>
</tr>
<tr>
<td>75-79</td>
<td>55</td>
</tr>
<tr>
<td>80+</td>
<td>132</td>
</tr>
<tr>
<td>OVER-WEIGHT</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td>97</td>
</tr>
<tr>
<td>25-29</td>
<td>127</td>
</tr>
<tr>
<td>30-34</td>
<td>126</td>
</tr>
<tr>
<td>35-39</td>
<td>111</td>
</tr>
<tr>
<td>40-44</td>
<td>131</td>
</tr>
<tr>
<td>45-49</td>
<td>92</td>
</tr>
<tr>
<td>50-54</td>
<td>99</td>
</tr>
<tr>
<td>55-59</td>
<td>70</td>
</tr>
<tr>
<td>60-64</td>
<td>131</td>
</tr>
<tr>
<td>65-69</td>
<td>102</td>
</tr>
<tr>
<td>70-74</td>
<td>90</td>
</tr>
<tr>
<td>75-79</td>
<td>66</td>
</tr>
<tr>
<td>80+</td>
<td>107</td>
</tr>
<tr>
<td>OBSE</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td>102</td>
</tr>
<tr>
<td>25-29</td>
<td>86</td>
</tr>
<tr>
<td>30-34</td>
<td>124</td>
</tr>
<tr>
<td>35-39</td>
<td>112</td>
</tr>
<tr>
<td>40-44</td>
<td>127</td>
</tr>
<tr>
<td>45-49</td>
<td>95</td>
</tr>
<tr>
<td>50-54</td>
<td>96</td>
</tr>
<tr>
<td>55-59</td>
<td>67</td>
</tr>
<tr>
<td>60-64</td>
<td>119</td>
</tr>
<tr>
<td>65-69</td>
<td>104</td>
</tr>
<tr>
<td>70-74</td>
<td>80</td>
</tr>
<tr>
<td>75-79</td>
<td>54</td>
</tr>
<tr>
<td>80+</td>
<td>41</td>
</tr>
</tbody>
</table>
5.4.2 Sorting Data File by Variables of Research Interest

The next step was to sort data according given the interest in estimating BP profiles of non-diabetic normal-weight, over-weight and obese persons. Here, the key variables are SBP, DBP, HbA1c, age, degree of excess weight and current hypertension or diabetes medications. BP and HbA1c levels are used globally for the diagnosis of hypertension and diabetes, and separately as risk markers for the development of complications. NHANES data include measurements for BP and HbA1c, which are taken manually by a trained operator according to a standard protocol.

We defined the study sample as non-diabetic subjects who were not taking diabetes medications at the time of the survey, with HbA1c levels of less than 7%, the target value assigned by the American Diabetes Association (Pani, et al., 2008). To limit our regression analysis to adults, we excluded subjects who were younger than 20 years old. We also excluded people who were taking hypertension medications that are designed to force BP levels down to the target values.

5.4.3 Regression Modeling (Age-Associated Physiological Trend)

The next major step was to model physiological changes associated with aging. We conducted polynomial regression to explain the changing patterns of SBP, DBP, and HbA1c with age. The sorted samples in the study were divided into 5-year age groups in order to provide a reasonable sample size for each group, as shown in Table 5.2. Outcomes were the mean values for physiological variables in every age group, analyzed by age. All analyses considered differential weights of selection and the complex sample design, and sampling weights were adjusted for
unequal probabilities of selection resulting from nonresponse and planned oversampling of certain subgroups.

The next step was to create the regression model to track the changing patterns in SBP, DBP, and HbA1c with age, referred to in Table 5.2. As demonstrated in the literature, all organ systems are affected by aging (Boss & Seegmiller, 1981; Lazarus & Harridge, 2010). Patterns in the physiological processes of interest (i.e., SBP, DBP, HbA1c) evaluated in the NHANES (1999-2008) also have been proven to be associated with aging. We examined the difference in mean SBP, DBP, and HbA1c by age group using analysis of variance (ANOVA). Trend tests were carried out using multiple regression analysis to explain changes in SBP, DBP, and HbA1c with increasing age. The regression coefficients were evaluated by the mean levels of the physiological variables with age and the regression models were fitted as polynomial equations satisfying 95% of adjusted $R^2$ as follows:

<table>
<thead>
<tr>
<th></th>
<th>SBP</th>
<th>DBP</th>
<th>HbA1c</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NORMAL-WEIGHT</strong></td>
<td>$SBP = 109 - 0.1163 \times Age + 0.00675 \times Age^2$</td>
<td>$DBP = 54.64 + 0.3641 \times Age + 0.00611 \times Age^2 - 0.00017 \times Age^3$</td>
<td>$HbA1c = 5.048 - 0.008384 \times Age + 0.000415 \times Age^2 - 0.000003 \times Age^3$</td>
</tr>
<tr>
<td><strong>OVER-WEIGHT</strong></td>
<td>$SBP = 112.7 - 0.0655 \times Age + 0.005483 \times Age^2$</td>
<td>$DBP = 38.04 + 1.608 \times Age - 0.01929 \times Age^2 + 0.000041 \times Age^3$</td>
<td>$HbA1c = 5.146 - 0.01161 \times Age + 0.000504 \times Age^2 - 0.000003 \times Age^3$</td>
</tr>
<tr>
<td><strong>OBESE</strong></td>
<td>$SBP = 107.8 + 0.3091 \times Age + 0.001165 \times Age^2$</td>
<td>$DBP = 21.39 + 2.834 \times Age - 0.04428 \times Age^2 + 0.000193 \times Age^3$</td>
<td>$HbA1c = 5.290 - 0.01722 \times Age + 0.000728 \times Age^2 - 0.000005 \times Age^3$</td>
</tr>
</tbody>
</table>
Table 5.2. Means and standard errors (SE) of SBP, DBP, and HbA1c by age and obesity categories in the non-diabetic population. (Source: NHANES 1999-2008)

<table>
<thead>
<tr>
<th>AGE</th>
<th>SBP (mmHg)</th>
<th></th>
<th>DBP (mmHg)</th>
<th></th>
<th>HbA1c (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
</tr>
<tr>
<td>20-24</td>
<td>111.02</td>
<td>0.47</td>
<td>64.78</td>
<td>0.42</td>
<td>5.05</td>
</tr>
<tr>
<td>25-29</td>
<td>111.64</td>
<td>0.48</td>
<td>66.94</td>
<td>0.43</td>
<td>5.06</td>
</tr>
<tr>
<td>30-34</td>
<td>111.48</td>
<td>0.58</td>
<td>68.43</td>
<td>0.39</td>
<td>5.10</td>
</tr>
<tr>
<td>35-39</td>
<td>111.31</td>
<td>0.63</td>
<td>69.28</td>
<td>0.54</td>
<td>5.16</td>
</tr>
<tr>
<td>40-44</td>
<td>113.78</td>
<td>0.72</td>
<td>71.33</td>
<td>0.65</td>
<td>5.21</td>
</tr>
<tr>
<td>45-49</td>
<td>119.99</td>
<td>0.94</td>
<td>75.06</td>
<td>0.57</td>
<td>5.26</td>
</tr>
<tr>
<td>50-54</td>
<td>121.74</td>
<td>1.04</td>
<td>74.94</td>
<td>0.67</td>
<td>5.35</td>
</tr>
<tr>
<td>55-59</td>
<td>124.51</td>
<td>1.48</td>
<td>73.76</td>
<td>0.64</td>
<td>5.43</td>
</tr>
<tr>
<td>60-64</td>
<td>131.22</td>
<td>1.22</td>
<td>73.04</td>
<td>0.70</td>
<td>5.45</td>
</tr>
<tr>
<td>65-69</td>
<td>130.78</td>
<td>1.28</td>
<td>71.78</td>
<td>0.76</td>
<td>5.47</td>
</tr>
<tr>
<td>70-74</td>
<td>136.22</td>
<td>1.67</td>
<td>67.29</td>
<td>0.66</td>
<td>5.48</td>
</tr>
<tr>
<td>75-79</td>
<td>137.28</td>
<td>1.32</td>
<td>66.32</td>
<td>1.12</td>
<td>5.57</td>
</tr>
<tr>
<td>80+</td>
<td>145.70</td>
<td>1.42</td>
<td>62.01</td>
<td>0.73</td>
<td>5.58</td>
</tr>
<tr>
<td>20-24</td>
<td>114.83</td>
<td>0.62</td>
<td>64.98</td>
<td>0.65</td>
<td>5.09</td>
</tr>
<tr>
<td>25-29</td>
<td>115.36</td>
<td>0.55</td>
<td>68.03</td>
<td>0.58</td>
<td>5.13</td>
</tr>
<tr>
<td>30-34</td>
<td>115.56</td>
<td>0.51</td>
<td>71.24</td>
<td>0.57</td>
<td>5.20</td>
</tr>
<tr>
<td>35-39</td>
<td>116.87</td>
<td>0.57</td>
<td>71.88</td>
<td>0.49</td>
<td>5.23</td>
</tr>
<tr>
<td>40-44</td>
<td>117.83</td>
<td>0.63</td>
<td>74.09</td>
<td>0.56</td>
<td>5.28</td>
</tr>
<tr>
<td>45-49</td>
<td>121.43</td>
<td>0.97</td>
<td>75.88</td>
<td>0.59</td>
<td>5.33</td>
</tr>
<tr>
<td>50-54</td>
<td>125.38</td>
<td>0.70</td>
<td>76.32</td>
<td>0.45</td>
<td>5.43</td>
</tr>
<tr>
<td>55-59</td>
<td>126.39</td>
<td>0.89</td>
<td>75.74</td>
<td>0.60</td>
<td>5.49</td>
</tr>
<tr>
<td>60-64</td>
<td>132.05</td>
<td>1.12</td>
<td>72.75</td>
<td>0.69</td>
<td>5.53</td>
</tr>
<tr>
<td>65-69</td>
<td>132.33</td>
<td>1.06</td>
<td>71.18</td>
<td>0.75</td>
<td>5.61</td>
</tr>
<tr>
<td>70-74</td>
<td>137.48</td>
<td>1.15</td>
<td>68.77</td>
<td>0.79</td>
<td>5.60</td>
</tr>
<tr>
<td>75-79</td>
<td>138.59</td>
<td>1.75</td>
<td>64.96</td>
<td>0.80</td>
<td>5.68</td>
</tr>
<tr>
<td>80+</td>
<td>144.11</td>
<td>1.27</td>
<td>63.82</td>
<td>0.65</td>
<td>5.67</td>
</tr>
<tr>
<td>20-24</td>
<td>115.96</td>
<td>0.79</td>
<td>65.60</td>
<td>0.67</td>
<td>5.22</td>
</tr>
<tr>
<td>25-29</td>
<td>117.02</td>
<td>0.67</td>
<td>67.84</td>
<td>0.63</td>
<td>5.24</td>
</tr>
<tr>
<td>30-34</td>
<td>117.89</td>
<td>0.60</td>
<td>72.04</td>
<td>0.58</td>
<td>5.31</td>
</tr>
<tr>
<td>35-39</td>
<td>120.14</td>
<td>0.64</td>
<td>76.09</td>
<td>0.59</td>
<td>5.37</td>
</tr>
<tr>
<td>40-44</td>
<td>122.85</td>
<td>0.69</td>
<td>77.17</td>
<td>0.58</td>
<td>5.39</td>
</tr>
<tr>
<td>45-49</td>
<td>125.21</td>
<td>0.85</td>
<td>77.35</td>
<td>0.53</td>
<td>5.55</td>
</tr>
<tr>
<td>50-54</td>
<td>127.01</td>
<td>1.02</td>
<td>76.59</td>
<td>0.45</td>
<td>5.62</td>
</tr>
<tr>
<td>55-59</td>
<td>129.41</td>
<td>1.21</td>
<td>74.40</td>
<td>0.66</td>
<td>5.70</td>
</tr>
<tr>
<td>60-64</td>
<td>132.31</td>
<td>0.90</td>
<td>72.30</td>
<td>0.61</td>
<td>5.66</td>
</tr>
<tr>
<td>65-69</td>
<td>134.58</td>
<td>1.43</td>
<td>70.74</td>
<td>0.83</td>
<td>5.79</td>
</tr>
<tr>
<td>70-74</td>
<td>135.65</td>
<td>1.25</td>
<td>66.96</td>
<td>0.89</td>
<td>5.77</td>
</tr>
<tr>
<td>75-79</td>
<td>136.61</td>
<td>1.39</td>
<td>66.51</td>
<td>0.92</td>
<td>5.74</td>
</tr>
<tr>
<td>80+</td>
<td>141.98</td>
<td>1.69</td>
<td>62.23</td>
<td>0.96</td>
<td>5.78</td>
</tr>
</tbody>
</table>
Figure 5.2. Mean value of physiological variables (SBP, DBP, HbA1c) in the non-diabetic population
Age-specific physiological variables of non-diabetic participants by obesity classifications in NHANES

Figure 5.2 gives the regression graphs that show the relationships between age and variables
SBP, DBP, and HbA1c. These regression model outcomes are represented by the blue line. In
order to represent the degree of certainty for fit, we show 95% confidence intervals for predicted responses from the raw data with red dots.

These regression models and graphs representing the change patterns for SBP, DBP, and HbA1c with increasing age illustrate that the rise in SBP continues steadily throughout life with one inflection point in DBP and HbA1c. DBP rises until the age range of 50-60 and then tends to level off. There is a positive association between HbA1c and age groups until age 70, with a negative relationship thereafter.

5.4.4 Creating a BP Profile for One Person

The last step in PDE methodology is to estimate the patient profile. The main idea is to draw a profile estimate of one person, first by defining one person who satisfies the research objective conditions, then finding other analogous people in terms of physiological conditions, and then, yielding a profile estimate by combining the physiological variables of all selected people.

The main idea behind determining a profile of estimate of one person is finding analogous subjects in terms of physiological variables and then combining them. If there are specific conditions associated with the research objective such as age or range of physiological variables, a subject satisfying these conditions can be identified in the dataset and defined as the “base sample.”

Assuming there is a specific rate of aging for each physiological variable, the regression models can be adjusted by changing the value of the y-intercept in order to include the base sample. The revised regression model (which includes the base sample) can lead to finding other analogous people. Euclidean norm of residuals, which is known for the degree of best-fit
matching with the regression model, can be used to rank the degree of similarity to the base sample. The people with the lowest Euclidean norm of residuals can be regarded as having a high degree of similarity with the base sample. The number of people with lower Euclidean norm of residuals values can be adjusted based on research demands. The SBP and DBP levels of these selected people are adjusted based on the estimated levels of the base sample’s age. The final profile estimate for one person is obtained by bringing together all the physiological variables in order, beginning with lowest Euclidean norm of residuals.

As seen in Figure 5.3, age-associated patterns exist for SBP, DBP, and HbA1c. SBP and HbA1c increase continuously; DBP, however, rises linearly until a person is approximately in her 50s, and then tends to decrease later in life, as proven in numerous studies illustrating the relationship between SBP, DBP, and HbA1c and age (Burt, et al., 1995; Franklin, et al., 1997; Franklin, et al., 2009; Nuttall, 1999; Pani, et al., 2008). Even if the pattern of consistently increasing HbA1c levels are observed with advancing age, inflection points are observed when a person is in his or her 50s. Based on these observed changes in DBP and HbA1c levels, it is worth estimating BP profiles for people in their 50s. The PDE methodology will produce three cohort BP profiles by obesity classifications (normal-weight, over-weight, and obese people) as seen in Figure 5.3.

To create profile estimates of BP for one person, a base sample among the ages of 50 through 59 was selected and SBP, DBP and HbA1c levels were approximated as regression outcomes as seen in Figure 5.3. In order to obtain 75 BP measurements, the number of subjects required to create one profile is 25 including the base sample since the NHANES measures BP three times per person.
We found that the physiological variables SBP, DBP, and HbA1c have slopes based on age. Noting that there are no changes in the gradients of the regression models, we revised the regression models to include the SBP, DBP and HbA1c levels of the base sample by moving the intercept up or down along y-axis. Euclidean norm of residuals was used to find the other 24 subjects who were physiologically analogous to the base sample. The 24 people with the lowest Euclidean norm of residuals values were finally selected, and the BP levels of these 24 people were adjusted by estimating BP levels to the base sample’s age. Thus, a one-person profile estimate for BP was created. Repeating these PDE procedures, thirty-person BP profiles were created for three cohorts based on BMI classification, as shown in Figure 5.3.
5.5 Results

All of the SBP, DBP and HbA1c values were averaged for five-year population age groups, and repeated measures ANOVA was used to evaluate whether the age differences elicited significant changes in SBP, DBP and HbA1c, as seen in Table 5.3. The null hypothesis predicted equal changes in SBP, DBP and HbA1c among age groups. Considering a $p$-value of 0.05 to be significant, repeated-measures ANOVA yields that the hypothesis can be confidently rejected because the $F$-ratio for this hypothesis is very large with an associated $p$-value < 0.001. This implies that there are statistically significant differences for SBP, DBP and HbA1c among the age groups.

Table 5.3. SBP and DBP levels for normal-weight, over-weight and obese cohorts.

<table>
<thead>
<tr>
<th>Obesity Classification</th>
<th>Normal-weight</th>
<th>Over-weight</th>
<th>Obese</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBP (mmHg)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>120.96</td>
<td>125.00</td>
<td>128.44</td>
</tr>
<tr>
<td>SE</td>
<td>0.36</td>
<td>0.29</td>
<td>0.14</td>
</tr>
<tr>
<td>DBP (mmHg)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>72.98</td>
<td>74.94</td>
<td>78.20</td>
</tr>
<tr>
<td>SE</td>
<td>0.17</td>
<td>0.17</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Positive associations exist between BP and BMI, as established in several other studies including Neter, Stam, Kok, Grobbee, & Geleijnse, 2003 and Wilsgaard, Schirmer, & Arnesen, 2000. Therefore, it is reasonable to find this relationship using PDE methodology in the three cohort groups. Significant differences are observed in SBP and DBP levels with respect to different body obesity classifications (normal-weight, over-weight, and obese), as shown in Table 5.3. People with more excess body weight had higher SBP levels than normal-weight people. Similar findings also exist for DBP levels, with excess body weight being positively
associated with DBP level. Statistically significant differences in SBP and DBP levels based on degree of obesity are revealed in the repeated measures ANOVA analysis with a 95% confidence interval and a $p$-value $< 0.001$.

### 5.6 Discussion

Chronic disease cannot generally be prevented or cured with medication, but must be managed indefinitely. Considering the long-term latency and relapse conditions associated with chronic disease, it is necessary to track the change of relevant disease markers. However, most surveys that can be accessed by general researchers contain cross-sectional data, which do not track sequential properties over time. To address this gap, we propose the PDE methodology. The PDE methodology attempts to create patient profile estimates based on health outcomes with temporal properties using population-based health survey data.

To facilitate understanding of the PDE methodology, each step was explained using a case study where we created patient profile estimates based on BP health outcome data. The ability to produce patient profile estimates for BP is important, since the diagnosis, control and treatment of hypertension are defined by BP levels. This case study uses BP data for three cohorts based on BMI classification: normal-weight, over-weight, and obese. The resultant patient profiles show significant differences in BP levels among the three cohorts. Increasing patterns of SBP and DBP levels with BMI are observed, which is consistent with prior research results showing a positive association between BP and BMI.

The strength of the PDE methodology lies in its applicability to a wide range of health conditions. In this study, we created BP profiles based on obesity classification. We found that
excess body weight is associated with elevated BP levels. Other than obesity, there are other risk factors related to lifestyle choices that could be the main cause of elevated BP. Smoking, alcohol consumption, poor diet, and lack of exercise are common factors associated with hypertension. These other behavior factors also can be considered to define the cohorts and can be applied using the PDE methodologies to produce the BP levels. The BP profile estimates can then be used to analyze resultant impacts on hypertension.

The PDE methodology can be applied to other disease markers such as glycol-hemoglobin (HbA1c) and cholesterol level. If the age-associated patterns are proven in the regression models in the prior literature, all the physiological variables can be estimated using the PDE methodology. In this case study, HbA1c levels were considered in order to evaluate age-associated patterns and to find people with similar physiological variable values. However, HbA1c profiles can be estimated using SBP and DBP levels.

The patient profile estimates generated using the PDE methodology can be essential references for researchers who are exploring new hypotheses prior to incurring the expense of implementation. First, synthetic profile estimates can be used as potential evidence to prove the importance of variability in physiological variables. Most individual health diagnoses are made based on the underlying value or mean of a disease marker even though such measurements are variable. This implies that we would expect to observe variability in the measurement over time; such variability could signal the future onset or progression of disease, or highlight the imprecision of clinical measurements.

The outcomes of the PDE methodology can be used as a validation monitoring tool. Statistical quality control (SQC) and time-series models are common monitoring techniques for
observing the range or values of expected future outcomes in industrial settings. Currently, these methodologies are applied to monitor the health of individual patients as well as cohorts with chronic disease. Patient profile estimates for disease markers can be useful data sources demonstrating the appropriateness of the monitoring tools.

We must note the limitations of our study. First, the age-associated patterns are evaluated using polynomial regression to produce BP profile estimates among normal-weight, over-weight, and obese people. The polynomial regression models explain the patterns of SBP, DBP and HbA1c levels with age to satisfactory levels (95%) of $R^2$ which is the indicator determining the degree of fit with the regression models. The regression models can use different and more complex forms of equations when the physiological variables or cohorts of research interest change.

The PDE methodology relies on the availability and timelines of population-based survey data. We used NHANES (1999-2008) data, combining five two-year datasets, in order to estimate BP profiles. The PDE methodology can be applied with other datasets as well. The scale of study can be reduced by considering one two-year NHANES (2007-2008) dataset instead. Likewise, NHANES data after 2007-2008 can be added to the 10 years of data we used to provide more stable profile estimates.

The main contribution of the PDE methodology is to provide repeated measures of patient clinical profile estimates based on health outcomes in population-based survey data, which do not support consecutive measurements over time. To reduce failure rates in clinical research studies, researchers often conduct analytical studies using alternative data in order to validate their research interests. Most accessible datasets are population-based cross-sectional data
without temporal properties. The PDE methodology can resolve the limitations associated with longitudinal data by drawing profile estimates based on health outcomes in the cross-sectional data. The resulting profiles can be essential references for clinical studies prior to implementation.
CHAPTER 6

CONTRIBUTIONS AND FUTURE RESEARCH

Successful investigation of the health monitoring system brings together statistical and analytical tools for managing health related data that enable sensitive and timely detection of disease changes. In light of health monitoring concerns, this research contributes to preparedness and response planning for unexpected disease processes in both the immediate and long terms. This chapter summarizes the contributions of this research to health monitoring systems and suggests areas for future research.

6.1 Contributions

The dissertation provides enhanced healthcare data monitoring with SQC methods which can contribute to delivering accurate, timely and up-to-date information to patients, care-teams and relevant organizations in order for them to determine appropriate interventions. Several potential research contributions emerged from this research, which are summarized below.

6.1.1 Integrated Health Monitoring System with SQC, Regression Modeling, and Time Series Modeling

The nature of health data creates barriers to the direct use of traditional SQC methods, in health monitoring applications of SQC methods. SQC methods have basic underlying assumptions: the monitoring data are independently and normally distributed (Montgomery, 2009). However,
health data are more unlikely to satisfy the basic assumptions, unlike applications in industrial settings where there is a more controlled environment, a target/nominal value, and close approximation to the normal distribution. For example, we monitor the weekly ILI incidence data and weekly drug sales data corresponding to the ILI symptoms for syndromic surveillance. Evaluating the statistical characteristics of these syndromic data, they tend to be non-stationary (i.e., reflect a seasonal pattern), auto-correlated and non-normally distributed, which is in violation of the basic SQC assumptions. In addition, blood pressure measurements used in chronic disease monitoring tend to exhibit high autocorrelation which influences individual–level health outcomes. It is likely that different approaches to the data sources are required because of their respective characteristics in health monitoring applications.

In line with the nature of statistical characteristics in health data and the underlying assumptions of SQC, we suggested ways to integrate SQC methods and statistical forecasting methods of time series and regression models into the health monitoring endeavor. A trigonometric regression model was used to stabilize data flow when the data stream drifted away from having a fixed mean. Then two combined time series models based on analogue methodology and vector ARIMA were considered in order to forecast an upcoming event and produce the residuals following general SQC assumptions. The integrated health monitoring system can create opportunities of customized approach in order to deal with variable statistical characteristics inherent in health data which lead to provide more accurate, sensitive and timely information about disease changes between/ within patients.
6.1.2 Improvement of Health Monitoring SQC Methods: MCP and MSS-EWLR

Industrial SQC control charts often focus on the detection of step shifts in mean levels with a narrower focus on a specific aspect of the manufacturing process. The measurement systems and data collection processes associated with manufacturing applications are often validated and standardized up front.

However, health monitoring applications are often characterized by a considerable amount of uncontrollable variations, including somewhat irregular seasonal effects and distributed data collection processes, where errors and delays are more likely to occur (Woodall 2006, 2008; Tsui, et al., 2009). Because the baseline data can be contaminated by unexplained noise spikes that may adversely affect the performance of detection methods, it is important to strengthen the detection ability of existing surveillance methods.

Also, many types of outbreak or change patterns are associated with diseases. Research efforts are required to focus on developing robust monitoring methods that detect these various patterns in disease changes. In line with the characteristics of health data, we suggest two specific SQC monitoring methods: multivariate change point (MCP) and multivariate self-starting exponentially weighted likelihood ratio (MSS-EWLR) charts. These two tools provide quite robust and sensitive monitoring performance for various types of shifts (mean and/or covariance shifts) without long sequences of historical data, which allows complex and multifaceted variations inherent in health data to be monitored. In both charts, a statistical alarm signals an aberration in the monitored health data that may indicate the onset of potential disease progression or dispersion. The proposed health surveillance system provides advance warning given by the monitoring tools, such that the time delay between detecting the first symptoms and
implementing treatment activities could be shortened, enabling more effective treatment and reducing negative impact.

**Multivariate statistical process monitoring**

In health monitoring applications, process variables and biological characteristics are correlated with one other, so applying the univariate method multiple times to each individual variable can provide limited performance in detecting process changes. Faster detection of changing disease patterns and trends could be achieved by the use of multiple data streams and multivariate monitoring methods that consider variables jointly.

Multivariate methods are designed to treat all data simultaneously and extract information regarding the directionality of a process variation in terms of how the variables behave relative to each other. When several events occur in a process, they affect not only the magnitude of the variable variation, but also their relationships to each other (the direction of variation). It is often difficult to understand the magnitude of each process variable under such conditions because the signal-to-noise ratio is low. Multivariate methods can extract confirming information based on observations of many variables and can reduce noise levels through averaging by considering the covariance matrix that assigns a degree of relevance to each of the variables. Thus, the suggested MSS-EWLR and MCP charts, with their multivariate approach can explain the correlations among all variables, as well as detect abnormal signals more rapidly and accurately than the univariate approach.
**Monitoring variability in health surveillance**

In traditional applications of SQC tools in health monitoring systems, observations are monitored to determine if they are out of control based on the mean values alone even if variability in the disease process can also provide valuable information for managing healthcare. In order to allow for the recognition of advance symptoms and changes in disease, it is necessary to monitor variability in addition to the mean in disease process.

In multivariate monitoring settings, monitoring only the mean shift cannot provide a sensitive monitoring signal. Generally, clinical data are correlated with each other, so monitoring these biological characteristics separately without accounting for the correlations between them can limit abnormality detection. Also, changes in a process mean vector can be masked by unsuspected changes in the covariance structure or by a sudden change in the correlation between two variables. Thus, it is essential to develop robust SQC charts that can monitor the mean vector and covariance matrix simultaneously. To deal with this challenge in health monitoring, two SQC tools – the MSS-EWLR and MCP charts – were introduced which enable changes to be detected in the mean vector and/or covariance structure.

**Monitoring sequence of data without Phase I parameter information**

The general charting process relies on known or assumed known in-control (baseline) true parameters such as the mean vector or covariance matrix and uses the assumed true parameter values to set the control limits. In health monitoring applications, parameter information during in-control processes corresponding to baseline or stable conditions are rarely known. As an alternative, the parameters can be estimated using a large data sequence that is believed to
represent an in-control process corresponding to the baseline of disease. However, it is unrealistic to gather this series. Even if a dataset could be prepared to observe data patterns and estimate parameters for an in-control process, it could not be guaranteed that the obtained parameters would describe it adequately. To deal with this limitation, the MSS-EWLR and MCP charts both enable monitoring without parameter information \textit{a priori}.

6.1.3 Creating Synthetic Clinical Profiles from Cross-Sectional Survey

It is important to have clinical data that explain disease processes within/between patients in order to obtain accurate and timely information associated with disease changes. While up-to-date information on population health and behaviors may be publicly available in national surveys, there are significant access restrictions for repeated measurements of clinical data of an individual patient because of privacy issues.

In response to ongoing need for time-series data of health outcomes, we proposed the profile data estimation (PDE) methodology to transform cross-sectional (national survey) data into repeated measurement estimates that describes individual patient profile of health outcomes. The profile estimates throughout the PDE methodology can provide essential references for the researcher to do the pilot studies to find validation of new research as well as to relieve the burdens of validating new researches by reducing time and data collection cost.

We applied this methodology to national health and nutrition examination survey (NHANES) data in order to obtain blood pressure profiles of non-diabetic normal-weight, over-weight, and obese persons. The profiles of three cohort groups by obesity classifications validated that excess
body weight can elevate blood pressure levels. Also, synthetic profile estimates can be used to demonstrate the performance of proposed health monitoring SQC methods.

### 6.1.4 Multiple Streams of Health Data: Clinical and Non-clinical Data

Existing health monitoring systems mainly focus on one single source of clinical data. However, a single data source is not enough to identify, forecast, or explain illness trends, so non-clinical data sets may be used to complement existing practices in health surveillance (Burkom, Murphy, Coberly, & Hurt-Mullen, 2005; Rolka, et al., 2007). In line with this, we consider not only multiple data sources but also non-clinical data in concert with clinical data in order to track the movements of illness which enable faster and more certain detection of changes in disease.

**Multiple data sources**

A key to effective disease management is to understand movements in illness patterns, including direction and relative speed. A single data source is not enough to identify, forecast or explain illness trends. Considering multiple data sources that track the movements of illness enables faster and more certain detection than does a single data source; this is because the latter does not capture individual responses to abnormal symptoms. In order to improve our ability to detect diseases at their onset, the use of multiple data streams is necessary. Simply put, it is difficult to explain all the factors in disease transition (progression/diversion) with just a single data source. In syndromic surveillance example, we showed that multivariate monitoring with 22 regions of weekly ILI incidence data provided earlier indication of the outbreak than those provided in the univariate monitoring of national weekly ILI incidence data.
Non-clinical data sources

Perceptive considerations of non-clinical data have the potential to widen the scope of traditional surveillance by providing earlier indications of disease progression. Clinical data, such as laboratory test outcomes and vital signs, are used to diagnose health states as well as illness trends that encompass a disease’s diversion and progression. However, it is difficult to explain individual responses to illness symptoms using clinical data. Non-clinical data can help detect illness patterns at an early stage, although it is an indirect approach to explaining illness trends. In Chapter 3, we observed that OTC drug sales can be used to track responses to initial ILI symptoms rather than ILI incidence data for syndromic surveillance purposes. An analysis of OTC drug sales data verified that non-clinical data can provide an indirect, but earlier signal of an influenza outbreak.

6.2 Future Research

To facilitate disease prevention and control and better healthcare, there are growing interests in leveraging healthcare data. This will not advance the field of health monitoring, but also amplify the potential of SQC based methodologies that can be extended to other emerging areas in healthcare system. Specific directions of future work are discussed below.

6.2.1 Data Collection Efforts for Health Monitoring: Non-clinical Data

The purpose of data acquisition is to model the underlying patterns of disease processes that are likely to mask signals regarding disease progression/dispersion. To achieve this goal, non-
clinical data can be used to track the behavior of sick people before they visit a doctor. They have the potential to supply highly relevant information that can supplement traditional health monitoring.

Syndromic surveillance generally used clinical data related to patient symptoms (e.g., cough, fever, shortness of breath) during the early phase of illness. However, there are potential benefits to using non-clinical data which can capture behaviors of a population prior to visits to medical facilities; this has been highlighted in Shmueli & Burkom, 2010; Yan, Chen, & Zeng, 2009. Generally, activities responding to ILI symptoms in the prodromal phase of disease (non-clinical data) such as call volume to telephone triage advice lines, absences from school or work and over-the-counter drug sales may be self-treatment approaches used before or in lieu of visiting a doctor.

In order to evaluate this idea, Chapter 3 examined a clinical dataset of weekly ILI incidence for 22 regions and 20 classes of non-clinical data (i.e., OTC and prescribed weekly medication sales) to determine when a disease outbreak occurred. This example of syndromic surveillance proves that non-clinical data used in concert with clinical data can provide an earlier signal of an unexpected change in a disease pattern and can also act as an alternate source of information for understanding a disease’s dispersion. Non-clinical data can explain these population behaviors in the prodromal phase, creating the potential to provide an indirect but earlier signal of an outbreak.

Moving forward, non-clinical data can be used as an alternative source for identifying disease progression in chronic disease monitoring. Traditional approaches in chronic disease monitoring, generally focused on understanding and monitoring laboratory outcomes and vital signs in order
to catch up disease recurrence or progression. Changes in a patient’s clinical variables, such as blood glucose and blood pressure, may be due to changes in the patient’s underlying condition or biological processes, measurement error, or random variation. All efforts to monitor these clinical data streams are aimed at detecting disease process changes in the clinical variable and distinguishing them from background noise in order to facilitate appropriate clinical decision making.

Recently, a number of studies have been conducted to elucidate the impact of psychological and behavior changes as well as somatic states, which have already been shown to influence disease progression because traditional research approaches based on self-report and laboratory outcomes have limitations to assess complex and temporarily dynamic process of disease progression (Barnett, Spence, Manuck, & Jennings, 1997; Cohen & Herbert, 1996; Cohen, Tyrrell, & Smith, 1991; Kannel, et al., 1999; Krantz & McCeney, 2002; Spiegel & Giese-Davis, 2003). In attempt to articulate the associations of these behavior and psychological attributes to the health outcomes as well as to provide proactive care to patients, behavior medicine research and ecological momentary assessment (EMA) studies have been adopted. For example, non-clinical data related to behaviors like drinking alcohol, smoking and poor diet that can be triggered by severe stress can be considered surrogate sources to track the health outcomes for chronic disease. To validate the use of non-clinical data in chronic disease management, further studies using non-clinical data as well as clinical data are required.
6.2.2 Development of SQC Methods following Other Statistical Distributions

One of the general assumptions of the SQC method is that process data follows a normal distribution. The health data we used in syndromic surveillance and blood pressure monitoring followed a normal distribution after the preconditioning and forecasting procedures were followed, which enabled us to use a general SQC charting methodology. In other health monitoring applications, the health data may have different probability distributions such as binomial or Poisson distributions.

Generally, disease incidence rates occurring within a specific geographic area and subpopulation are common observations in public health surveillance. When health practitioners consider the number of incidences \( (p) \) corresponding to the disease of interest among \( n \) hospital visits in a given interval, the incidence rate \( \left( \frac{p}{n} \right) \) can follow a binomial distribution. If the expected number of occurrence in a fixed interval is given, the probability for the incidence rate can follow a Poisson distribution.

In line with various probability distributions of health data, it is necessary to develop SQC methods following binomial or Poisson distributions, since most SQC methodologies are defined based on a normal distribution. Therefore, advanced charting methods defined in the binomial or Poisson distribution must be created in order to extend use of SQC methodologies for health monitoring purposes.

6.2.3 Integration of SQC and Variability Research for Accurate Health Diagnosis

It is important to monitor physical conditions when a patient is diagnosed. Typically, physical conditions are identified with mean levels, which are compared with a predefined and classified
numerical range. Even if the true underlying value of a disease marker in an individual patient is stable, we would expect to observe measurement variability over time. Information about variability in the disease process can also provide valuable information for analyzing health conditions.

As proven in recent research, elevated BP and visit-to-visit variability have been found to significantly increase the risk of CVD mortality and morbidity in patients with type II diabetes as well as non-diabetic people (Hsieh, et al., 2012; Rothwell, et al., 2010). Previously, most of guidelines for managing BP relied on average BP level. However, recent research outcomes highlight BP variability as an important therapeutic component of chronic disease management.

Another example also illustrates the importance of variability information. Assume we have two patients who are diagnosed as being in a state of pre-hypertension. The deviation scope from expected value per measurement can be quite different, even when the overall means for these two patients are at the same level. If these two persons are diagnosed based solely on mean values, the only obtainable information is the diagnosis of pre-hypertension. In fact, a patient with high variability could have a significantly higher probability of being pre-diabetic than one with low variability in hypertension, according to Gupta (2008).

As seen in the two examples, variability in biological characteristics such as blood pressure, cholesterol, or glycol-hemoglobin can provide important information about disease changes. A high or low level of variability in observed health data may imply a prodromal signal for the presence of an expected chronic disease, as well as an unstable state of health which, if undetected, can worsen. Considering the benefits of SQC methods that enable us to distinguish special causes from background noise, SQC potentially can explain the changes of variability
that occur with time. Therefore, further research incorporating variability analysis and SQC in diagnosing health status is necessary.

6.2.4 Research on the Effect of Interventions in Chronic Disease Surveillance

One of the big issues in chronic disease surveillance is identifying the risk factors and patterns associated with chronic disease and assessing the impact of intervention programs to see if risk factors increase or decrease over time. When risk factors such as smoking or obesity must be monitored and mitigated, public health practitioners can suggest intervention strategies aimed at improving overall health in the population. The incidence rate is observed in order to assess whether these risk factors increase or decrease over time. When a reduction in the number of case reports or events occurs over time, it implies that the intervention strategy is effective. If the changes are observed in a sequence of observations, there might be one or more inflection points over time period, implying that the intervention program is impacting the population.

If the objective of chronic disease surveillance is to detect changes associated with an intervention program, we can apply the suggested health monitoring system. Generally, changes in the chronic disease index can occur in a sustained way, so traditional SQC applications designed to monitor for persistent mean shifts are well suited to chronic disease surveillance. Change in disease processes can be represented by OOC signals, implying that an intervention strategy is influencing the population. The main factors and timing (change-point in a process) causing OOC signals can be analyzed statistically. These approach enables further consideration of the effectiveness of interventions strategies.
For example, every state implemented a series of tobacco control measures from 2002 to 2006 that included raising the average price of the product through taxes, prohibiting smoking in essentially all indoor workplaces, or implementing media campaigns aimed at discouraging smoking (CDC, 2007). The morbidity and mortality weekly report (2007) provided data on the relative difference in the proportion of smokers in the population in 2002 and 2006 in order to draw statistical conclusions about smoking trends and impacts of these intervention programs. A statistical assessment of the significance of the change can be made using a simple pairwise comparison of the proportions. If we use the suggested health monitoring framework in this example, we can potentially provide valuable information such as which intervention program is the most effective at reducing the smoking prevalence rate, as well as when the smoking control program became effective. As seen in this example, the suggested health monitoring framework can be applied to evaluate the impact of intervention programs in chronic disease surveillance. Further studies are required to validate the use of this framework.
REFERENCES


VITA

Min-Jung Kim was born in Seoul, the Republic of Korea. She received her B.S. degree in Mathematics and in Management from the Chung-Ang University, Seoul, Korea, in 2001, and M.S. and Ph.D. degrees in industrial engineering from the Pennsylvania State University, University Park, PA, in 2009 and 2012 respectively. She was with DKSH Korea Ltd., Seoul, Korea, from 2001 to 2005, working on market trend analysis and registration consulting in new drug and chemical development for pharmaceutical companies. Her research interests lie in the area of health monitoring systems, especially statistical quality engineering tools to improve the quality of healthcare delivery. The research presented in this dissertation involves one published paper and four papers to be submitted to refereed journals. She is a member of American Society for Quality (ASQ).