RISK-ADJUSTED CUMULATIVE SUM CONTROL CHARTING PROCEDURES AND STANDARDIZED MORTALITY RATIOS

TANG XU

NATIONAL UNIVERSITY OF SINGAPORE
2013
RISK-ADJUSTED CUMULATIVE SUM CONTROL CHARTING PROCEDURES AND STANDARDIZED MORTALITY RATIOS

TANG XU
(SUPERVISED BY A/P GAN FAN FATT)

A THESIS SUBMITTED FOR THE DEGREE OF PHILOSOPHY OF DOCTORATE
NATIONAL UNIVERSITY OF SINGAPORE
2013
Acknowledgement

I would like to thank my supervisor, Associate Professor Gan Fah Fatt for his meritorious guidance, encouragement, time and endless patience. I also would like to thank Dr Zhang Lingyun for obtaining the cardiac data set from Steiner, to be used for illustration of my procedures in the thesis. Next, I will thank my family for their sincere love and support. Finally, special thanks to my friends who had helped me.
Summary

The contents of the thesis are organized into 3 chapters. The risk-adjusted cumulative sum charting procedures for multiple responses are developed in Chapter 1. The standardized mortality ratios are developed in Chapter 2, followed by a comparison of standardized mortality ratios in Chapter 3. Although Chapters 2 and 3 are related, all the chapters are organized in such a way that they can be read independently. This is to facilitate the submission of these chapters for publication in journals.
## CONTENTS

**Acknowledgement**

**Summary**

### Chapter 1  Risk-Adjusted Cumulative Sum Charting Procedures for Multiple Responses

Abstract 1

§1 Introduction 2

§2 Cumulative Sum Charts for Multi-Responses 4

§3 Average Run Length of Cumulative Sum Charts 21

§4 Sensitivity of the Cumulative Sum Chart with Respect to the Risk Distribution 25

§5 Applications 28

§6 Comparison with CUSUM Chart Based on Binary Outcomes 44

§7 Conclusion 55

Appendix A: CDF and PDF of \(W(Y, S)\) 57

Appendix B: Collocation Method 59

Appendix C: R Codes for Binary Outcome Case 61

Appendix D: R Codes for 3-outcome Case 73

References 86

### Chapter 2  Standardized Mortality Ratio

Abstract 89

§1 Introduction 90

§2 Statistical Properties of SMR 94

§3 Comparison of Confidence Intervals of SMR 97

§4 Power of the 4 Test Procedures 106

§5 Effect of Risk Distributions on SMR 113

§6 Real Data Analysis 117
Chapter 1. Risk-Adjusted Cumulative Sum Charting Procedures for Multiple Responses

Abstract

A risk-adjusted cumulative sum (CUSUM) charting procedure has been developed in the literature for monitoring the performance of a surgeon or a group of surgeons. This chart assumes that a surgical operation results in one of two outcomes: survival and dead. Such a classification is naive because a patient who has a full recovery is considered the same as another patient who survived but remained bed-ridden for life. For a patient who survives an operation, there are different grades of recovery. It thus makes more sense to consider a risk-adjusted CUSUM charting procedure based on more than two outcomes. In this thesis, we develop such a chart and study its performance. The advantages of using more than two outcomes are demonstrated with real data sets. Our chart is shown to be a valuable tool for analyzing and comparing the performances of surgeons.

KEY WORDS: Cardiac operations; Collocation method; Euroscores; Odds ratio; Parsonnet scores; Proportional odds logistic regression; Quality monitoring.
§1. Introduction

Monitoring quality of medical practice is becoming a key component in improving efficiency in health care. The necessity of formally monitoring surgical outcomes has been brought to the forefront due to some critical cases (see Treasure et al. 1997). For instance, in 1999, the UK General Medical Council found three doctors possibly guilty of professional misconduct over the quality of heart surgeries performed. This misconduct led to 29 mortalities out of 53 children who were operated at the Bristol Royal Infirmary. After that, the UK National Institute of Clinical Excellence was established to monitor medical professional conduct in hospitals. In such cases, the quickness of detection of deterioration in surgical operations is critical because it will promote prompt investigation of the causes and possibly lead to a reduction in mortalities.

Monitoring quality of medical procedures is different from monitoring of processes in an industrial setting where the raw material fed into a manufacturing process is assumed to be homogeneous. Patients in a hospital differ from each other notably in terms of their health conditions. A patient's health condition like age, blood pressure, existence of certain disease like diabetes, morbid obesity et al. is commonly summarized using a score. For example, the Parsonnet score (Parsonnet et al., 1989) is used to summarize the risk of death resulting from a cardiac operation. The heterogeneity of patients must be taken into account in any monitoring procedure to prevent making any misleading inferences. Any procedure that takes the patients' risks into account is known as a risk-adjusted procedure. Another widely used score is the Euroscore which was developed by Roques et al. (1999) by fitting a logistic regression model based on 19030 cardiac surgeries. The Euroscore uses a patient's characteristics like age, gender, serum creatinine, systolic pulmonary pressure, neurological dys-function etc to estimate the mortality risk of the patient undergoing a cardiac surgery.

In order to take the risk of a patient into account, Lovegrove et al. (1997,
1999) and Poloniecki et al. (1998) suggested estimating the probability of death $x_i$ from the Parsonnet score and then plotting the difference between $x_i$ and the surgical outcome $y_i$ (1 if a patient dies or 0 if the patient survives within 30 days) cumulatively. They termed this chart a variable life-adjusted display (VLAD) and cumulative risk-adjusted mortality (CRAM) chart respectively. This chart is intuitive, easy to understand and it accounts for the risk of a patient by producing a risk-adjusted score $x_i - y_i$ for monitoring. The score compares directly the patient’s risk and the outcome of the operation. Treasure et al. (2004) presented some convincing examples using the VLAD that showed both improvement and deterioration in performances of some surgeons. The main criticism of this chart is the lack of a proper signaling rule. Although Poloniecki et al. (1998) proposed certain control limits for signaling but these limits are not directly interpretable in terms of run length performance. A signaling rule provides an objective and quantitative way for assessing the points plotted and determining appropriate times for taking action. Sherlaw-Johnson (2005) mapped the control limits of the risk-adjusted cumulative sum chart onto the VLAD but the resulting signaling rule is complicated because the control limits change with inclusion of data from every new surgical operation.

The CUSUM chart was initially developed by Page (1954) for monitoring manufacturing processes. Moustakides (1986) showed that the CUSUM chart is optimal in terms of run length criterion. The CUSUM chart was first proposed by Williams et al. (1992) to monitor surgical performances. De Leval et al. (1994) and Steiner et al. (1999) consider a problem of monitoring outcomes in paediatrics cardiac surgery. However, Steiner et al. (1999) did not take the mortality risk into account because he claimed that patient characteristics did not have a significant effect on the mortality rate. Later, Steiner et al. (2000) proposed a risk-adjusted CUSUM chart based on testing the odds ratio that a patient dies, as a way to account for the patient’s risk. The risk-adjusted CUSUM chart and VLAD are developed using two different approaches and are viewed as two different charts (Spiegelhalter, 2003; Rogers et
al., 2004; Sherlaw-Johnson, 2005; Woodall, 2006), but Gan et al. (2012) showed that the CUSUM chart and the CUSUM chart based on the VLAD’s monitoring statistic are in fact the same.

Up till now, research is based only on binary outcomes resulting from a surgery. However, the outcome from a surgery can be more meaningfully classified into more than 2 categories. For example, the outcomes of a cardiac surgery can be categorized into the following categories: (1) death, (2) return to operating room, (3) postoperative stroke, (4) mediastinitis, (5) postoperative atrial fibrillation, (6) full recovery (Shortell et al., 2000). Other than death, the other outcomes represent different grades of a ‘successful’ operation. It is clear that outcomes (2) to (5) should not be considered the same as outcome (6). It makes good sense to have more than 2 outcomes to represent the outcome of a surgery. In Section 2, we develop a CUSUM charting procedure for monitoring a surgical process with more than two outcomes. A proportional odds logistic regression model is used to estimate the probabilities of various surgical outcomes. Conditions are then established for reasonable risk-adjusted log likelihood ratio score. An accurate method of computing the average run length (ARL) of the CUSUM is described in Section 3. The sensitivity of the CUSUM chart with respect to the risk distribution is examined in Section 4. We show the effectiveness of the chart using 2 simulated data sets in Section 5. The chart is then used to study and compare the performances of 7 surgeons based on a real data set. A comparison between CUSUM charts based on 2 and 3 outcomes is given in Section 6. This comparison shows the advantage of using 3 outcomes instead of 2 outcomes. Finally, a conclusion is given in Section 7.

§2. Cumulative Sum Charts for Multi-Responses

The CUSUM charting procedure was first developed for the manufacturing industries. Page (1954) introduced the CUSUM chart and demonstrated its ability to detect small but persistent shifts. Suppose $X_n$ is the monitoring statistic based on the $n$th sample obtained. Let the process parameter of interest be denoted by $\theta$
and the probability density function (pdf) of $X_n$ be denoted by $f(x_n; \theta)$. In order to detect a change in $\theta$, let $\theta_0$ and $\theta_A$ be the in-control and out-of-control parameters respectively. The CUSUM chart is obtained by plotting

$$C_n = \max(0, C_{n-1} + W_n),$$

against the sample number $n$ where $C_0 = u, 0 \leq u < h$ and $W_n$ is the log likelihood ratio statistic

$$W_n = \log \left\{ \frac{f(X_n; \theta_A)}{f(X_n; \theta_0)} \right\}. \quad (2.2)$$

The chart signals when $C_n$ exceeds the control limit $h$ and a signal indicates that the CUSUM chart has accumulated sufficient evidence to indicate a change in the parameter.

The performance of the CUSUM chart is commonly measured by the average run length (ARL). A small in-control ARL may result in many false signals but it is more sensitive to changes in the process parameter. The ARL of the CUSUM chart under the null hypothesis $H_0$ is analogous to the probability of type I error of a hypothesis test, while the ARL under the alternative hypothesis $H_A$ is analogous to the power of the test. Moustakides (1986) showed that the CUSUM chart is optimal in the sense that it gives the smallest ARL under $H_A$ among all charts with the same in-control ARL.

When the CUSUM chart is used to monitor surgical performances as described in the introduction, the underlying mortality risk of a patient will have to be taken into account along with the outcome of a surgery. Suppose a patient is to undergo a cardiac surgery. The patient’s conditions like age, blood pressure, existence of certain disease like diabetes, morbid obesity etc. will be determined and this information can be summarized as a risk score, $S$. The risk score is a real number and it is a measure of the mortality risk of a patient undergoing a cardiac surgery. The outcome of a cardiac surgery is usually assessed after 30 days and it can be represented by an integer variable $Y$ which takes a value from 0 to $J$. Let $Y = J$ when a
patient dies, $Y = 0$ when a patient has a fully recovery, and $Y = 1, 2, \ldots, J - 1$ are used to represent various states of partial recovery, with a smaller number associated with a better state of recovery.

Conditional on a patient’s risk score $S = s$, the outcome $Y$ follows a multiple response distribution

$$P(Y = k|S = s) = \pi_k(s), \; k = 0, 1, \ldots, J.$$ 

We consider Surgeon $A$ to be uniformly better than surgeon $B$ if

$$P(Y_A \leq k|S = s) \geq P(Y_B \leq k|S = s)$$

holds for all $k$ and all $s$. This means that a patient’s outcome is more likely to be in a better state when the patient is operated on by surgeon $A$. A patient’s risk score can be used to estimate the probability mass function of $Y$ by using a multiple response logistic model which defines the cumulative probabilities as

$$P(Y \leq k|S = s) = \pi_0(s) + \cdots + \pi_k(s), \; k = 0, \ldots, J - 1.$$ 

The cumulative logits are then given as

$$\text{logit}[P(Y \leq k|S = s)] = \log \left[ \frac{P(Y \leq k|S = s)}{1 - P(Y \leq k|S = s)} \right] = \log \left[ \frac{\pi_0(s) + \cdots + \pi_k(s)}{\pi_{k+1}(s) + \cdots + \pi_J(s)} \right].$$

The proportional odds logistic regression model (McCullagh, 1980) uses all the $J$ cumulative logits simultaneously as follows

$$\text{logit}[P(Y \leq k|S = s)] = \alpha_k + \beta s, \; k = 0, \ldots, J - 1. \quad (2.3)$$

The model assumes that each cumulative logit has its own intercept $\alpha_k$ and share the same slope effect $\beta$. The assumption that all the logit surfaces are parallel is commonly known as the proportional odds assumption. For a proportional odds
logistic regression model to be valid, this assumption must be tested. The standard test is a score test (Brant, 1990; Peterson and Harrell, 1990; Bender and Crouven, 1997). One attractive property of the proportional odds logistic model is that only the sign inversion of the regression parameters occur when the variable Y codes are inverted. Another attractive feature of this model is that the odds ratio for each predictor is taken to be constant across all possible collapsing of the response variable. The proportional odds logistic regression model is the most popular method for ordinal data (see Hastie et al., 1989; Brant, 1990; Woodward et al., 1995; Bender and Grouven, 1997; Hamid and Stephenson, 2006; Scott et al., 2010; Das and Rahman, 2011; Citko et al., 2012). The $\alpha_k$ is increasing in $k$ because the probability $P(Y \leq k | S = s)$ increases in $k$ for all $s$ and the logit is an increasing function of this probability. Note that $\beta$ is negative for our application because the risk score represents the risk which has a negative effect on the cumulative probability $P(Y \leq k | S = s)$.

In monitoring a surgical process, it is logical to take a patient’s conditions into account. Let the pdf of the risk score of a patient be represented by $f(s)$. The joint density of $(S, Y)$ is then given as $f(s, y) = \pi_y(s)f(s)$, $y = 0, \ldots, J$. In addition, we assume $H_0 : (\pi_0(s), \ldots, \pi_J(s)) = (\pi^0_0(s), \ldots, \pi^0_J(s))$ and $H_A : (\pi_0(s), \ldots, \pi_J(s)) = (\pi^A_0(s), \ldots, \pi^A_J(s))$, under the null and alternative hypotheses respectively. The risk distribution $f(s)$ is assumed to be the same for both hypotheses. The CUSUM chart for monitoring $Y$ is then given by equation (2.1) where $C_0 = u$ and $W_n$ is the sequential log likelihood ratio statistic

$$ W_n = \log \left\{ \frac{f_A(S_n, Y_n)}{f_0(S_n, Y_n)} \right\}. $$

The statistic $W_n$ is risk-adjusted by taking $S_n$ into account. It follows that the joint pdf’s under the null and alternative hypotheses are given by $f_0(s_n, y_n) = \pi^0_{y_n}(s_n)f(s_n)$ and $f_A(s_n, y_n) = \pi^A_{y_n}(s_n)f(s_n)$ respectively. Hence,

$$ W_n = \log \left\{ \frac{f_A(S_n, Y_n)}{f_0(S_n, Y_n)} \right\} = \log(\pi^A_{Y_n}(S_n)/\pi^0_{Y_n}(S_n)). \quad (2.4) $$
Note that \( W_n \) does not contain \( f(\cdot) \) because we assume the risk distribution to be the same for both hypotheses. In practice, this means that the patients’ risk profiles remain unchanged.

For the binary case where \( J = 1 \), the null and alternative hypotheses can be represented by \( H_0 : (\pi_0(s), \pi_1(s)) = (\pi^0_0(s), \pi^0_1(s)) \) and \( H_A : (\pi_0(s), \pi_1(s)) = (\pi^A_0(s), \pi^A_1(s)) \). The statistic \( W_n \) is then reduced to

\[
W_n = \begin{cases} 
\log(\pi^0_0(S_n)/\pi^0_1(S_n)) & \text{if } Y_n = 0; \\
\log(\pi^A_0(S_n)/\pi^A_1(S_n)) & \text{if } Y_n = 1.
\end{cases}
\]

We can define the performance of a surgeon in terms of the risk of mortality. Suppose the estimated risk of mortality of the \( n \)th patient, calculated using a logistic model based on the past performances of a group of surgeons, is represented by \( p_n(s) \) or \( p_n \).

The performance of a surgeon can be characterized by the odds ratio of mortality \( Q \) as

\[
Q = \frac{\pi^1_1(s)/(1 - \pi^0_1(s))}{p_n/(1 - p_n)}.
\]

(2.5)

where \( \pi^1_1(s) \) is the risk of mortality of the \( n \)th patient operated on by the surgeon.

We can define the null and alternative hypotheses in terms of odds ratio of mortality

\[
H_0 : \text{odds ratio } = Q_0, \\
H_A : \text{odds ratio } = Q_A.
\]

The conditional probabilities of mortality can then be simplified as

\[
(\pi^0_0(s), \pi^0_1(s)) = \left( \frac{1 - p_n}{1 - p_n + Q_0 p_n}, \frac{Q_0 p_n}{1 - p_n + Q_0 p_n} \right),
\]

and

\[
(\pi^A_0(s), \pi^A_1(s)) = \left( \frac{1 - p_n}{1 - p_n + Q_A p_n}, \frac{Q_A p_n}{1 - p_n + Q_A p_n} \right).
\]

Therefore,

\[
W_n = \begin{cases} 
\log\left[ \frac{1 - p_n + Q_0 p_n}{1 - p_n + Q_A p_n} \right] & \text{if } Y_n = 0; \\
\log\left[ \frac{(1 - p_n + Q_0 p_n) Q_A}{(1 - p_n + Q_A p_n) Q_0} \right] & \text{if } Y_n = 1.
\end{cases}
\]
The statistic $W_n$ for the binary case was first proposed by Steiner et al. (2000).

For the binary case, the performance of a surgeon is defined using the odds ratio $Q$ given in equation (2.5). If the same approach is used to define the performance in a multi-response case, we will obtain

$$\frac{\pi^*_k(s)}{1 - \pi^*_k(s)} = Q_k \frac{\pi_k(s)}{1 - \pi_k(s)}, \quad k = 0, \ldots, J - 1.$$  

However, based on this definition, the odds ratios $Q_k, k = 0, \ldots, J - 1$ must satisfy some stringent and complicated conditions to guarantee that the probabilities $\pi^*_k(s), k = 0, \ldots, J$ to be in $[0, 1]$. This definition is associated with the logistic model based on binary outcomes but it is not appropriate for the multi-response case. For the proportional odds logistic regression model, a more natural way of defining performance is to use the one based on cumulative probabilities

$$\frac{\sum_{i=0}^{k} \pi^*_i(s)}{1 - \sum_{i=0}^{k} \pi^*_i(s)} = R_k \frac{\sum_{i=0}^{k} \pi_i(s)}{1 - \sum_{i=0}^{k} \pi_i(s)}, \quad k = 0, \ldots, J - 1,$$  

(2.6)

where $R_k$ is the odds ratio based on cumulative probabilities of recovery. In order for $\pi^*_k(s), k = 0, \ldots, J$ to be in $[0, 1]$, the odds ratios must satisfy the condition

$$\alpha_0 + \log(R_0) \leq \alpha_1 + \log(R_1) \leq \cdots \leq \alpha_{J-1} + \log(R_{J-1}).$$

This definition of odds ratio is consistent with our earlier definition that Surgeon A is uniformly better than surgeon $B$ if $P(Y_A \leq k|S = s) \geq P(Y_B \leq k|S = s)$ holds for all $j$ and all $s$. For Steiner et al. (2000), note that the odds ratio is defined using the probability of death rather than recovery.
Figure 2.1. Plots of $W(y,s)$ against the Parsonnet score $s$ for $y = 0, 1, 2$ when $J = 2$ and $R_0 = 0.5, R_1 = 0.4$.

Figure 2.2. Plots of $W(y,s)$ against the Parsonnet score $s$ for $y = 0, 1, 2$ when $J = 2$ and $R_0 = 2, R_1 = 2.5$.
Figure 2.3. Plots of $W(y, s)$ against the Parsonnet score $s$ for $y = 0, 1, 2$ when $J = 2$ and $R_0 = R_1 = 0.5$.

Figure 2.4. Plots of $W(y, s)$ against the Parsonnet score $s$ for $y = 0, 1, 2$ when $J = 2$ and $R_0 = R_1 = 2$. 
In practice, it is reasonable to assume the odds ratios of cumulative probabilities to be 1 under the null hypothesis, that is $R_0 = \ldots = R_{J-1} = 1$. In other words, we assume the performance under the null hypothesis is characterized by the fitted proportional odds logistic regression model. According to equation (2.6), for detecting an improvement, we can set all the $R_k$’s to take values greater than 1. Similarly, for detecting a degradation, we can set all the $R_k$’s to take values less than 1. Once an alternative hypothesis is chosen, the monitoring statistic $W(Y, S)$ can then be calculated using equation (2.4).

The risk-adjusted score $W(Y, S)$ can be interpreted as a penalty score in a CUSUM chart for detecting degradation. It is reasonable that the penalty score decreases as the risk score increases given any of outcomes $Y = k, k = 0, \ldots, J$, because a surgeon should be given a lower penalty score for a higher-risk patient given the same outcome. It is also reasonable to have a higher penalty score for a less desirable outcome given the risk score $s$. Similarly, the risk-adjusted score $W(Y, S)$ can be interpreted as a reward score in a CUSUM chart for detecting improvement. It is reasonable that the reward score increases as the risk score increases given any of outcomes $Y = k, k = 0, \ldots, J$, because a surgeon should be awarded a higher reward score for a higher-risk patient given the same outcome. It is also reasonable to have a lower reward score for a more severe outcome given the risk score $s$.

Consider $R_0 = 0.5$ and $R_1 = 0.4$ as the alternative hypothesis for detecting a degradation and $R_0 = 2$ and $R_1 = 2.5$ as the alternative hypothesis for detecting an improvement. *Figures* 2.1 and 2.2 show the plots of $W(y, s)$ against $s$ for detecting degradation and improvement respectively. Note that the lines associated with partial recoveries and death cross when the Parsonnet score is 50.5 in *Figure* 2.1. This means that for a patient with a Parsonnet score $S$ less than 50.5, the risk-adjusted score (penalty score) $W(y, s)$ is higher if the patient makes a partial recovery rather than a death. In *Figure* 2.2, the lines associated with full and partial recoveries cross when the Parsonnet score is 54.5. This means that for a
patient with a Parsonnet score greater than 54.5, the risk-adjusted score (reward score) $W(y, s)$ is higher if the patient makes a partial recovery rather than a full recovery. In other words, the risk-adjusted score $W(y, s)$ is not reasonable given the choice of the odds ratios $R_0$ and $R_1$ in Figure 2.1 because the risk-adjusted score $W(2, s)$ is not uniformly greater than $W(1, s)$ for all $s$. Similarly, the risk-adjusted score $W(y, s)$ is not reasonable as shown in Figure 2.2 because the risk-adjusted score $W(0, s)$ is not uniformly greater than $W(1, s)$ for all $s$. It is then necessary to investigate appropriate choice of odds ratios for the outcome $Y$ to be properly risk-adjusted.

In order for the outcome $Y$ to be properly risk-adjusted, the risk-adjusted score $W(Y, S)$ must satisfy the following 2 properties:

Property 1: Conditional on $Y = k$, $W(k, s)$ is a monotonic decreasing function of $s$ when detecting degradation. Conditional on $Y = k$, $W(k, s)$ is a monotonic increasing function of $s$ when detecting improvement.

Property 2: For detecting degradation, $W(0, s) \leq W(1, s) \leq \ldots \leq W(J, s)$ for all $s$. For detecting improvement, $W(0, s) \geq W(1, s) \geq \ldots \geq W(J, s)$ for all $s$.

The conditions for Properties 1 and 2 to hold are given in Theorems 1 and 2 respectively.

**Theorem 1.** Assume equations (2.3), (2.4) and (2.6) hold. (i) If $R_k \leq 1, k = 0, \ldots, J - 1$, $W(y, s)$ is a decreasing function of $s$ given $y$. (ii) If $R_k \geq 1, k = 0, \ldots, J - 1$, $W(y, s)$ is a increasing function of $s$ given $y$.

**Proof.** From the proportional odds logistic regression model (2.3),

$$\text{logit}[P(Y \leq k | S = s)] = \alpha_k + \beta s, k = 0, \ldots, J - 1,$$
we can obtain the conditional probability

\[
\pi_k(s) = P(Y \leq k \mid S = s) - P(Y \leq k - 1 \mid S = s)
= \frac{\exp(\alpha_k + \beta s)}{1 + \exp(\alpha_k + \beta s)} - \frac{\exp(\alpha_{k-1} + \beta s)}{1 + \exp(\alpha_{k-1} + \beta s)}
= \frac{\exp(\beta s)[\exp(\alpha_k) - \exp(\alpha_{k-1})]}{[1 + \exp(\alpha_k + \beta s)][1 + \exp(\alpha_{k-1} + \beta s)]},
\]

where \( k = 0, \ldots, J \), \( \alpha_{-1} = -\infty \) and \( \alpha_J = \infty \). From equations (2.3) and (2.6), we have

\[
\log \left( \frac{\sum_{i=0}^{k} \pi_i^A(s)}{[1 - \sum_{i=0}^{k} \pi_i^A(s)]} \right) = \log(R_k) + \alpha_k + \beta s. \tag{2.7}
\]

From the proportional odds logistic regression model (2.7), we can obtain

\[
\pi_k^A(s) = \frac{\exp(\beta s)[\exp(\alpha_k + \log(R_k)) - \exp(\alpha_{k-1} + \log(R_{k-1}))]}{[1 + \exp(\alpha_k + \log(R_k) + \beta s)][1 + \exp(\alpha_{k-1} + \log(R_{k-1}) + \beta s)]}
\]

where \( k = 0, \ldots, J \), and \( R_{-1} = R_J = 1 \). Hence, equation (2.4) yields

\[
W(k, s) = \log[\pi_k^A(s)/\pi_k(s)] = D_k + \log(1 + \exp(\alpha_k + \beta s)) + \log(1 + \exp(\alpha_{k-1} + \beta s)) - \log(1 + \exp(\alpha_k + \log(R_k) + \beta s)) - \log[1 + \exp(\alpha_{k-1} + \log(R_{k-1}) + \beta s)],
\]

where

\[
D_k = \log[\exp(\alpha_k + \log(R_k)) - \exp(\alpha_{k-1} + \log(R_{k-1}))] - \log[\exp(\alpha_k) - \exp(\alpha_{k-1})].
\]

Taking the first derivative with respect to \( s \), we obtain

\[
W'(k, s) = \beta \left[ \frac{\exp(\alpha_k + \beta s)}{1 + \exp(\alpha_k + \beta s)} - \frac{\exp(\alpha_{k-1} + \beta s)}{1 + \exp(\alpha_{k-1} + \beta s)} - \frac{\exp(\alpha_{k-1} + \log(R_k) + \beta s)}{1 + \exp(\alpha_{k-1} + \log(R_k) + \beta s)} + \frac{\exp(\alpha_k + \log(R_k) + \beta s)}{1 + \exp(\alpha_k + \log(R_k) + \beta s)} \right]
= \beta \left[ \frac{1}{1 + \exp(\alpha_k + \log(R_k) + \beta s)} - \frac{1}{1 + \exp(\alpha_{k-1} + \log(R_{k-1}) + \beta s)} \right]
= \beta E.
\]
where
\[ E = \frac{1}{1 + \exp(\alpha_k + \log(R_k) + \beta s)} - \frac{1}{1 + \exp(\alpha_k + \beta s)} + \frac{1}{1 + \exp(\alpha_{k-1} + \log(R_{k-1}) + \beta s)} - \frac{1}{1 + \exp(\alpha_{k-1} + \beta s)}. \]

A condition for \( E \geq 0 \) is \( R_k \leq 1, k = 0, \ldots, J - 1 \). Similarly, a condition for \( E \leq 0 \) is \( R_k \geq 1, k = 0, \ldots, J - 1 \). In addition, note that \( \beta < 0 \) from earlier discussion. Thus, \( W'(k, s) \leq 0 \) if \( R_k \leq 1, k = 0, \ldots, J - 1 \) and \( W'(k, s) \geq 0 \) if \( R_k \geq 1, k = 0, \ldots, J - 1 \). This completes the proof.

\[ \text{Theorem 2.} \quad \text{Assume equations (2.3), (2.4) and (2.6) hold. (i) A necessary and sufficient condition for } W(y, s) \text{ to be an increasing function of } y \text{ given } s \text{ is } R_0 = R_1 = \cdots = R_{J-1} = R \leq 1. \text{ (ii) A necessary and sufficient condition for that } W(y, s) \text{ to be a decreasing function of } y \text{ given } s \text{ is } R_0 = R_1 = \cdots = R_{J-1} = R \geq 1. \]

\[ \text{Proof of (i).} \]

To prove necessity, assume \( W(y, s) \) is an increasing function of \( y \) given \( s \): \( W(0, s) \leq W(1, s) \leq \cdots \leq W(J, s) \). Equivalently, equation (2.4) yields
\[ \frac{\pi_0^A(s)}{\pi_0(s)} \leq \frac{\pi_1^A(s)}{\pi_1(s)} \leq \cdots \leq \frac{\pi_J^A(s)}{\pi_J(s)}. \]

Then, we have
\[ \frac{\pi_0^A(s)}{\pi_0(s)} \leq \frac{\pi_0^A(s) + \pi_1^A(s)}{\pi_0(s) + \pi_1(s)} \leq \cdots \leq \frac{\sum_{i=0}^{J-1} \pi_i^A(s)}{\sum_{i=0}^{J-1} \pi_i(s)} \leq \frac{\sum_{i=0}^{J} \pi_i^A(s)}{\sum_{i=0}^{J} \pi_i(s)} = 1, \tag{2.8} \]
\[ \frac{\pi_J^A(s)}{\pi_J(s)} \geq \frac{\pi_J^A(s) + \pi_J^A(s)}{\pi_J(s) + \pi_J(s)} \geq \cdots \geq \frac{\sum_{i=1}^{J} \pi_i^A(s)}{\sum_{i=1}^{J} \pi_i(s)} \geq \frac{\sum_{i=0}^{J} \pi_i^A(s)}{\sum_{i=0}^{J} \pi_i(s)} = 1. \tag{2.9} \]
Based on the odds ratio of cumulative probabilities defined in equation (2.6) we obtain

\[
\sum_{i=0}^{k} \pi_i^A(s) = \frac{R_k}{1 - \sum_{i=0}^{k} \pi_i(s) + R_k \sum_{i=0}^{k} \pi_i(s)}, \quad k = 0, \ldots, J - 1, \quad (2.10)
\]

\[
\sum_{i=k+1}^{J} \pi_i^A(s) = \frac{1}{1 - \sum_{i=0}^{k} \pi_i(s) + R_k \sum_{i=0}^{k} \pi_i(s)}, \quad k = 0, \ldots, J - 1. \quad (2.11)
\]

Substitute (2.10) into (2.8) and (2.11) into (2.9), we get

\[
\frac{R_0}{1 - \pi_0(s) + R_0 \pi_0(s)} \leq \frac{R_1}{1 - \sum_{i=0}^{1} \pi_i(s) + R_1 \sum_{i=0}^{1} \pi_i(s)} \leq \cdots \leq \frac{R_{J-1}}{1 - \sum_{i=0}^{J-1} \pi_i(s) + R_{J-1} \sum_{i=0}^{J-1} \pi_i(s)} \leq 1 \quad (2.12)
\]

and

\[
\frac{1}{1 - \sum_{i=0}^{J-1} \pi_i(s) + R_{J-1} \sum_{i=0}^{J-1} \pi_i(s)} \geq \frac{1}{1 - \sum_{i=0}^{J-2} \pi_i(s) + R_{J-2} \sum_{i=0}^{J-2} \pi_i(s)} \geq \cdots \geq \frac{1}{1 - \pi_0(s) + R_0 \pi_0(s)} \geq 1. \quad (2.13)
\]

From the definition of risk score, if \( s \to \infty \), \( \pi_J(s) \to 1 \), thus \( \sum_{i=0}^{k} \pi_i(s) \to 0 \) for \( k = 0, \ldots, J - 1 \) and we obtain the following from equation (2.12)

\[
R_0 \leq R_1 \leq \cdots \leq R_{J-1} \leq 1. \quad (2.14)
\]

Similarly, if \( s \to -\infty \), \( \pi_0(s) \to 1 \), thus \( \sum_{i=0}^{k} \pi_i(s) \to 1 \) for \( k = 0, \ldots, J - 1 \) and we obtain the following from equation (2.13)

\[
1 \leq 1/R_0 \leq 1/R_1 \leq \cdots \leq 1/R_{J-1}. \quad (2.15)
\]
(2.14) and (2.15) imply \( R_0 = R_1 = \cdots = R_{J-1} = R \leq 1 \).

To prove sufficiency, let \( R_0 = R_1 = \cdots = R_{J-1} = R < 1 \) and \( \Delta = \log(R) \).

Then, we have

\[
W(k, s) = \log(\frac{\pi_k^A(s)}{\pi_k(s)})
\]

\[
= \Delta + \log(1 + \exp(\alpha_k + \beta s)) + \log(1 + \exp(\alpha_{k-1} + \beta s))
- \log(1 + \exp(\alpha_k + \Delta + \beta s)) - \log(1 + \exp(\alpha_{k-1} + \Delta + \beta s)).
\]

Let \( g_k(\Delta) = W(k+1, s) - W(k, s) \). Then

\[
g_k(\Delta) = \log(1 + \exp(\alpha_{k+1} + \beta s)) - \log(1 + \exp(\alpha_{k+1} + \Delta + \beta s))
- \log(1 + \exp(\alpha_{k-1} + \beta s)) + \log(1 + \exp(\alpha_{k-1} + \Delta + \beta s)).
\]

It is clear that \( g_k(0) = 0 \). Take the first derivative of \( g_k(\Delta) \)

\[
g'_k(\Delta) = -\frac{\exp(\alpha_{k+1} + \Delta + \beta s)}{1 + \exp(\alpha_{k+1} + \Delta + \beta s)} + \frac{\exp(\alpha_{k-1} + \Delta + \beta s)}{1 + \exp(\alpha_{k-1} + \Delta + \beta s)}
= \frac{1}{1 + \exp(\alpha_{k+1} + \Delta + \beta s)} - \frac{1}{1 + \exp(\alpha_{k-1} + \Delta + \beta s)}.
\]

Note that \( g'_k(\Delta) < 0 \) because \( \alpha_k \) is increasing in \( k \). Hence, \( g_k(\Delta) < 0 \) if \( \Delta > 0 \) and \( g_k(\Delta) > 0 \) if \( \Delta < 0 \). Since \( R < 1 \), \( \Delta < 0 \) and \( g_k(\Delta) > 0 \). Thus, \( W(k+1, s) > W(k, s) \). In other words, \( W(Y, S) \) is an increasing function of \( Y \) conditional on \( S \). This proves the sufficiency. Result (ii) can be proved in a similar manner.

From Theorem 1, Property 1 of \( W(Y, S) \) is satisfied if the odds ratios of cumulative probabilities under the alternative hypothesis are set as \( R_k < 1, k = 0, \cdots, J-1 \) for detecting degradation and \( R_k > 1, k = 0, \cdots, J-1 \) for detecting improvement.

From Theorem 2, Property 2 holds if and only if the odds ratios of cumulative probabilities under the alternative hypothesis are chosen to be the same, that is \( R_1 = R_2 = \cdots = R_{J-1} \). Thus, in order for \( W(Y, S) \) to be reasonable and meaningful, we have to set the odds ratios of cumulative probabilities under the alternative hypothesis to be \( R_1 = R_2 = \cdots = R_{J-1} < 1 \) for detecting degradation and \( R_1 = R_2 = \cdots = R_{J-1} > 1 \) for detecting improvement.
If the condition on the odds ratios given in Theorem 2 holds, then additional properties of the risk-adjusted score \( W(y, s) \) can be derived and these are stated in Theorem 3.

**Theorem 3.**

Assume equations (2.3), (2.4) and (2.6) hold.

Case I: If \( R_0 = R_1 = \cdots = R_{J-1} = R < 1 \) (testing degradation), then the following are true.
1. If \( Y = 0 \), \( W(Y, s) < 0 \).
2. If \( Y = J \), \( W(Y, s) > 0 \).
3. \( W(0, s) \to 0 \) when \( s \to -\infty \), \( W(J, s) \to 0 \) when \( s \to \infty \).
4. For \( Y \in \{0, \cdots, J - 1\} \), \( W(Y, s) \to \log(R) \) when \( s \to \infty \).
5. For \( Y \in \{1, \cdots, J\} \), \( W(Y, s) \to -\log(R) \), when \( s \to -\infty \).

Case II: If \( R_0 = R_1 = \cdots = R_{J-1} = R > 1 \) (testing improvement), then the following are true.
1. If \( Y = 0 \), \( W(Y, s) > 0 \).
2. If \( Y = J \), \( W(Y, s) < 0 \).
3. \( W(0, s) \to 0 \) when \( s \to -\infty \), \( W(J, s) \to 0 \) when \( s \to \infty \).
4. For \( Y \in \{0, \cdots, J - 1\} \), \( W(Y, s) \to \log(R) \) when \( s \to \infty \).
5. For \( Y \in \{1, \cdots, J\} \), \( W(Y, s) \to -\log(R) \), when \( s \to -\infty \).

**Proof of Case (I).**

Let \( \Delta = \log(R) \). Note that \( \Delta < 0 \) because \( R < 1 \).

\[
W(k, s) = \log(\frac{\pi_k^A(s)}{\pi_k(s)}) = \Delta + \log(1 + \exp(\alpha_k + \beta s)) + \log(1 + \exp(\alpha_{k-1} + \beta s)) - \log(1 + \exp(\alpha_k + \Delta + \beta s)) - \log(1 + \exp(\alpha_{k-1} + \Delta + \beta s)).
\]

1. For \( Y = 0 \) and \( \alpha_{-1} = -\infty \), then

\[
W(0, s) = \Delta + \log(1 + \exp(\alpha_0 + \beta s)) - \log(1 + \exp(\alpha_0 + \Delta + \beta s)) = \log(1 + \exp(\alpha_0 + \beta s)) - \log(\exp(-\Delta) + \exp(\alpha_0 + \beta s)).
\]
Since $\Delta < 0$, $\exp(-\Delta) > 1$ and hence $W(0, s) < 0$.

2. For $Y = J$ and $\alpha_J = \infty$, then

$$W(J, s) = \log(1 + \exp(\alpha_{J-1} + \beta s)) - \log(1 + \exp(\alpha_{J-1} + \Delta + \beta s))$$

Since $\Delta < 0$, $W(J, s) > 0$.

3. This is clear from the functions of $W(0, s)$ and $W(J, s)$ given in parts (1) and (2).

4. and 5. For $Y = k \in \{1, \cdots, J - 1\}$,

$$W(k, s) = \Delta + \log(1 + \exp(\alpha_k + \beta s)) + \log(1 + \exp(\alpha_{k-1} + \beta s)) - \log(1 + \exp(\alpha_k + \Delta + \beta s)) - \log(1 + \exp(\alpha_{k-1} + \Delta + \beta s)).$$

Note that $\beta < 0$ for our logistic model.

Let $s \to \infty$, then $W(k, s) \to \Delta < 0$. Let $s \to -\infty$, then $W(k, s) \to -\Delta > 0$.

For $Y = 0$, from the function $W(0, s)$ obtained in part 1, let $s \to \infty$, then $W(0, s) \to \Delta$. For $Y = J$, from the function $W(J, s)$ obtained in part 2, we can show that

$$W(J, s) = \log(1 + \exp(\alpha_{J-1} + \beta s)) - \log(1 + \exp(\alpha_{J-1} + \Delta + \beta s))$$

$$= - \Delta + \log(1 + \exp(\alpha_{J-1} + \beta s)) - \log(\exp(-\Delta) + \exp(\alpha_{J-1} + \beta s)).$$

Let $s \to -\infty$, then $W(J, s) \to -\Delta > 0$. This completes the proof for Case (I). Case (II) can be proved in a similar manner.

For testing degradation, results 1 and 2 of Theorem 3 imply that the penalty score $W(0, s)$ is negative for full recovery and $W(J, s)$ is positive for death. The negative penalty score is equivalent to a reward when a patient makes a full recovery. This is reasonable because a full recovery is a desirable outcome and a death is an undesirable outcome. For partial recoveries, that is $k = 1, \cdots, J - 1$, results 4 and 5 imply that $W(k, s) > 0$ for $s$ less than some $s^*$ and $W(k, s) < 0$ for $s$ greater than some $s^*$. Thus, a patient with a risk $s$ less than $s^*$, the penalty is positive if the patient makes a partial recovery. In other words, a patient with a low risk and
has a partial recovery is not considered a desirable outcome. On the other hand, a patient with a risk $s$ greater than $s^*$, the penalty is negative. This means a patient with a high risk and has a partial recovery is considered a desirable outcome. This is again reasonable because if a high-risk patient makes even a partial recovery, this is considered a desirable outcome, whereas a low-risk patient who makes a partial recovery is not considered to be a desirable outcome. Similar interpretation can be obtained for testing improvement.

In a summary, the score $W(Y, S)$ can be interpreted as a penalty-reward score. Note that the score $W(0, s)$ is always a reward, and $W(J, s)$ is always a penalty. The score $W(k, s), k = 1, \ldots, J - 1$ can either be viewed as a penalty or a reward depending on the state of partial recovery and the risk of a patient. Furthermore, it can be seen from result 4 of Theorem 3 that for a very high risk patient, the reward given to a partial recovery is very close to that of a full recovery. This means that any state of partial recovery is considered as good as a full recovery for a very high risk patient. Similarly, result 5 of Theorem 3 implies that for a very low risk patient, the penalty given to a partial recovery is very similar to that of a patient who dies. This means that any state of partial recovery is considered as bad as dead for a very low risk patient. This is again a reasonable and desirable property of risk adjustment.

We will now summarize the procedure of constructing a CUSUM chart for monitoring the performance of a surgeon or a group of surgeons.

**Step 1.** Fit a proportional odds logistic regression model (2.3) using some past surgical data to estimate the probability of obtaining the outcome $k$, $\pi_k(s), k = 0, \ldots, J$ assuming the average performance of surgeons in the data set.

**Step 2.** The probability of obtaining the outcome $k$, $\pi^*_k(s), k = 0, \ldots, J$ assuming the odds ratios $R_0, R_1, \ldots, R_{J-1}$ for a surgeon or group of surgeon can be determined using equation (2.6).

**Step 3.** Set the null and alternative hypotheses as $H_0: R_0 = R_1 = \cdots = R_{J-1} = 1$
versus \( H_A : R_0 = R_1 = \cdots = R_{J-1} = R \). For a chart designed to detect a degradation, set \( R < 1 \). For improvement, set \( R > 1 \). The risk-adjusted score \( W_n \) is given in equation (2.4) where \( \pi_k^0(s) \) and \( \pi_k^A(s) \), \( k = 0, 1, \ldots, J \) are obtained from equation (2.6).

Step 4. The CUSUM chart is obtained by plotting equation (2.1).

Step 5. Determine the control limit \( h \) given an in-control ARL.

§3. Average Run Length of Cumulative Sum Charts

We have developed a CUSUM chart for monitoring surgical process with multiple outcomes in the last section. The odds ratio \( R \) specified in the null and alternative hypotheses and the control limit \( h \) determine the in-control ARL. In this section, we investigate a method of finding the ARL of the procedure based on the integral equation approach introduced by Page (1954).

Let \( L(u) \) denote the ARL of a CUSUM chart that starts with \( C_0 = u \), the integral equation for the ARL derived by Page (1954) is given as

\[
L(u) = 1 + L(0)F_W(-u) + \int_0^h L(x)f_W(x-u)dx,
\]

where \( F_W(\cdot) \) and \( f_W(\cdot) \) are the cumulative distribution function (cdf) and pdf of \( W(Y, S) \) respectively. The function \( L(u) \) can be approximated numerically by using the collocation method (Knoth, 2005). Gan et al. (2012) showed that the collocation method is sufficiently accurate for the case \( J = 1 \) (2 outcomes). We will examine the accuracy of this method for the case \( J = 2 \) (3 outcomes).

We consider a real data set of 6449 Parsonnet scores of cardiac patients and its distribution is shown in Figure 3.1 (Steiner et al. 2000). A real data set was used for the study so that any results obtained will be representative of real scenarios. Each patient was operated on by a surgeon from a group of surgeons. We set the outcome of a surgery to be \( Y = 2 \) if a patient dies within 30 days of the surgery, \( Y = 1 \) if there is partial recovery, and \( Y = 0 \) for full recovery. A proportional odds
logistic regression model (2.3) is first fitted using the data set as

\[
\log \left[ \frac{\pi_0(s)}{\pi_1(s) + \pi_2(s)} \right] = \alpha_0 + \beta s, \\
\log \left[ \frac{\pi_0(s) + \pi_1(s)}{\pi_2(s)} \right] = \alpha_1 + \beta s,
\]

where \(\alpha_0 = 3.057, \alpha_1 = 3.691\) and \(\beta = -0.078\). Note that the score test for the proportional odds assumption gives a \(p\)-value of 0.3606 which is not significant. It is thus feasible to use the proportional odds logistic regression model. This model will be used throughout the chapter. The probabilities of obtaining the 3 outcomes assuming the average performance of surgeons in the data set can be determined as

\[
\pi_0(s) = \frac{\exp(\alpha_0 + \beta s)}{1 + \exp(\alpha_0 + \beta s)}, \\
\pi_2(s) = \frac{1}{1 + \exp(\alpha_1 + \beta s)}, \\
\pi_1(s) = 1 - \pi_0(s) - \pi_2(s).
\]

The performance of a surgeon characterized by \(R_0\) and \(R_1\) can then be determined using equation (2.6). When we set the null and alternative hypotheses as \(H_0 : R_0 = R_1 = 1\) and \(H_A : R_0 = R_1 = R\) respectively, the form of \(W(Y, S)\) will then be determined. In the study to examine the accuracy of the collocation method in approximating the ARL, we assume that the distribution of \(S = 100\) follows a beta distribution, beta\((a = 1, b = 3)\). Suppose the performance of a surgeon is characterized by \(R_0^*\) and \(R_1^*\), then the distribution of \(Y\) given \(S\) can be determined. Finally, the pdf and cdf of \(W(Y, S)\) can be derived. The derivation of the pdf and cdf of \(W(Y, S)\) is given in the Appendix A. The pdf and cdf of \(W(Y, S)\) depend not only on the null and alternative hypotheses, but also on the actual performance of the surgeon being monitored and the underlying pdf of \(S\). Once the pdf and cdf of \(W(Y, S)\) are determined, we can then examine the accuracy of the collocation method in approximating the ARL function. The details are given in Appendix B.

For a CUSUM chart designed to detect degradation with \(R = 0.5\), Tables 3.1 and 3.2 show the in-control and out-of-control ARLs obtained using both collocation
and simulation methods. For detecting an improvement in performance with $R = 2$, Tables 3.3 and 3.4 show the in-control and out-of-control ARLs obtained using the collocation and simulation methods. It can be seen that the ARLs obtained using the two methods are consistent, thus the collocation method is sufficiently accurate for approximating the ARL.

![Parsonnet score frequency distribution](image)

**Figure 3.1.** The frequency distribution of Parsonnet scores of a group of 6449 patients who underwent cardiac operations.
Table 3.1. The in-control ARLs of the CUSUM chart for detecting degradation

<table>
<thead>
<tr>
<th>Control Limit h</th>
<th>Method</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Collocation</td>
<td>10.6</td>
<td>36.0</td>
<td>85.2</td>
<td>174.3</td>
<td>329.1</td>
<td>590.4</td>
<td>1037.2</td>
<td>1772.3</td>
</tr>
<tr>
<td></td>
<td>Simulation</td>
<td>10.7</td>
<td>36.1</td>
<td>85.4</td>
<td>174.4</td>
<td>328.3</td>
<td>593.7</td>
<td>1037.2</td>
<td>1770.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.32)</td>
<td>(0.25)</td>
<td>(0.23)</td>
<td>(0.44)</td>
<td>(0.64)</td>
<td>(1.12)</td>
<td>(1.73)</td>
</tr>
</tbody>
</table>

Table 3.2. The out-of-control ARLs of the CUSUM chart for detecting degradation when the true performance of a surgeon is characterized by \( R_0^* = R_1^* = 0.5 \)

<table>
<thead>
<tr>
<th>Control Limit h</th>
<th>Method</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Collocation</td>
<td>5.9</td>
<td>14.1</td>
<td>23.6</td>
<td>34.0</td>
<td>45.0</td>
<td>56.4</td>
<td>68.0</td>
<td>79.7</td>
</tr>
<tr>
<td></td>
<td>Simulation</td>
<td>5.9</td>
<td>14.1</td>
<td>23.7</td>
<td>33.9</td>
<td>45.0</td>
<td>56.7</td>
<td>68.3</td>
<td>79.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.05)</td>
<td>(0.11)</td>
<td>(0.18)</td>
<td>(0.24)</td>
<td>(0.31)</td>
<td>(0.38)</td>
<td>(0.43)</td>
<td>(0.47)</td>
</tr>
</tbody>
</table>

Table 3.3. The in-control ARLs of the CUSUM chart for detecting improvement

<table>
<thead>
<tr>
<th>Control Limit h</th>
<th>Method</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Collocation</td>
<td>12.7</td>
<td>40.9</td>
<td>96.7</td>
<td>198.1</td>
<td>374.5</td>
<td>675.2</td>
<td>1180.0</td>
<td>2020.6</td>
</tr>
<tr>
<td></td>
<td>Simulation</td>
<td>12.7</td>
<td>40.5</td>
<td>96.4</td>
<td>198.0</td>
<td>374.0</td>
<td>675.5</td>
<td>1179.2</td>
<td>2019.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.10)</td>
<td>(0.34)</td>
<td>(0.27)</td>
<td>(0.26)</td>
<td>(0.49)</td>
<td>(0.64)</td>
<td>(1.1)</td>
<td>(1.96)</td>
</tr>
</tbody>
</table>

Table 3.4. The out-of-control ARLs of the CUSUM chart for detecting improvement when the true performance of a surgeon is characterized by \( R_0^* = R_1^* = 2 \)

<table>
<thead>
<tr>
<th>Control Limit h</th>
<th>Method</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>2.5</th>
<th>3.0</th>
<th>3.5</th>
<th>4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Collocation</td>
<td>7.9</td>
<td>18.2</td>
<td>30.6</td>
<td>44.3</td>
<td>58.8</td>
<td>73.7</td>
<td>89.0</td>
<td>104.4</td>
</tr>
<tr>
<td></td>
<td>Simulation</td>
<td>7.9</td>
<td>18.1</td>
<td>30.9</td>
<td>44.6</td>
<td>58.1</td>
<td>73.2</td>
<td>89.5</td>
<td>104.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.05)</td>
<td>(0.12)</td>
<td>(0.21)</td>
<td>(0.28)</td>
<td>(0.36)</td>
<td>(0.44)</td>
<td>(0.52)</td>
<td>(0.59)</td>
</tr>
</tbody>
</table>
§4. Sensitivity of the Cumulative Sum Chart with Respect to the Risk Distribution

The distribution of $W(Y, S)$ and hence the ARL of the CUSUM chart is jointly determined by the distributions of $Y$ and $S$. From analyses of surgical data, it is found that the beta distribution provides reasonable approximations to the distribution of $S/100$ in many cases. Hence, in this section, we consider the beta distributions $\text{beta}(a, b)$ with $(a, b) = (1, 2), (1, 2.5), (1, 3), (1, 4), (1, 5), (2, 5)$ and $(5, 1)$ to investigate the sensitivity of the CUSUM chart with respect to the risk distribution. The pdf’s of $S$ corresponding to these beta distributions are displayed in Figure 4.1. The distributions of $S/100$ from surgical data for cardiac operations tend to follow the shapes of the beta distributions with $a = 1$. We also consider $\text{beta}(2, 5)$ which has substantially more elevated risk patients while $\text{beta}(5, 1)$ represents a distribution where most patients are of extremely high risk.

![Figure 4.1. Plots of pdf of $S$.](image)
We consider the odds ratios $R = 0.125, 0.25, 0.5, 0.75$ and $0.875$ for detecting degradation and $R = 1.125, 1.25, 1.5, 1.75, 1.875, 2.0$ for detecting improvement as our alternative hypotheses. The control limit $h$ is chosen such that the in-control ARL of the CUSUM chart is 100 when $S/100$ follows a beta distribution beta($a = 1, b = 3$). In order to study the effect of the risk distribution on the in-control ARL, we calculate the ARL of these CUSUM charts for other distributions. All the ARLs are calculated using the collocation method described in Appendix B.

Tables 4.1 and 4.2 display the in-control ARLs under different risk distributions. The ARL decreases at most about 6.7% when the distribution of $S/100$ changes from beta($1, 3$) to beta($1, 2$) in which there are more high risk patients. Also, the ARL increases at most about 22.2% when the distribution of $S/100$ changes from beta($1, 3$) to beta($1, 5$) in which there are more low risk patients. Although these distributions share similar shapes, the changes in the ARLs are significant. However, neither can one establishes that more high risk patients in the underlying risk distribution result in a shorter ARL nor more low risk patients result in a longer ARL. In addition, the ARL decreases at most about 16.8% when the distribution of $S/100$ changes from beta($1, 3$) to beta($2, 5$) in which there are more medium-risk patients. Furthermore, the ARL increases to at least 170.8% when the distribution of $S/100$ changes from beta($1, 3$) to beta($5, 1$) in which most of the patients are of extremely high risk. It is clear that the risk distribution does affect the ARL of a CUSUM chart. The density of $W$ after risk-adjustment of $Y$ is dependent on a given risk distribution $S$. Hence, the density of $W$ changes with the risk distribution. This explains the changes in the in-control ARL when the risk distribution changes. This also means that any conclusion drawn from a CUSUM charting procedure may not be meaningful if there is a substantial change in the risk distribution. It is thus necessary and important to monitor the underlying risk distribution as proposed by Loke and Gan (2012) to avoid making any incorrect conclusion.
Table 4.1: Sensitivity analysis: The in-control ARLs of the CUSUM charts for detecting degradation under different risk distribution. The control limit $h$ is chosen such that the in-control ARL of the CUSUM chart is 100 when $S/100$ follows a beta distribution beta(1,3).

<table>
<thead>
<tr>
<th>CUSUM Chart</th>
<th>$R = 0.875$</th>
<th>0.75</th>
<th>0.5</th>
<th>0.25</th>
<th>0.125</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S/100$</td>
<td>$h = 0.423$</td>
<td>0.829</td>
<td>1.606</td>
<td>2.380</td>
<td>2.826</td>
</tr>
<tr>
<td>beta(1,2)</td>
<td>95.9</td>
<td>96.6</td>
<td>98.7</td>
<td>101.5</td>
<td>105.0</td>
</tr>
<tr>
<td>beta(1,2.5)</td>
<td>97.1</td>
<td>97.4</td>
<td>98.4</td>
<td>99.8</td>
<td>101.4</td>
</tr>
<tr>
<td>beta(1,3)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>beta(1,4)</td>
<td>108.4</td>
<td>107.7</td>
<td>105.7</td>
<td>103.3</td>
<td>99.9</td>
</tr>
<tr>
<td>beta(1,5)</td>
<td>118.3</td>
<td>116.9</td>
<td>112.9</td>
<td>108.1</td>
<td>101.5</td>
</tr>
<tr>
<td>beta(2,5)</td>
<td>83.8</td>
<td>84.0</td>
<td>84.7</td>
<td>86.1</td>
<td>88.6</td>
</tr>
<tr>
<td>beta(5,1)</td>
<td>223.8</td>
<td>232.6</td>
<td>256.0</td>
<td>291.6</td>
<td>327.5</td>
</tr>
</tbody>
</table>

Table 4.2: Sensitivity analysis: The in-control ARLs of the CUSUM charts for detecting improvement under different risk distributions. The control limit $h$ is chosen such that the in-control ARL of the CUSUM chart is 100 when $S/100$ follows a beta distribution beta(1,3).

<table>
<thead>
<tr>
<th>$R$ and $h$</th>
<th>$R = 1.125$</th>
<th>1.25</th>
<th>1.5</th>
<th>1.75</th>
<th>1.875</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S/100$</td>
<td>$h = 0.373$</td>
<td>0.655</td>
<td>1.055</td>
<td>1.325</td>
<td>1.431</td>
<td>1.522</td>
</tr>
<tr>
<td>beta(1,2)</td>
<td>94.9</td>
<td>94.6</td>
<td>94.1</td>
<td>93.7</td>
<td>93.5</td>
<td>93.3</td>
</tr>
<tr>
<td>beta(1,2.5)</td>
<td>96.6</td>
<td>96.5</td>
<td>96.2</td>
<td>96.1</td>
<td>96.0</td>
<td>95.9</td>
</tr>
<tr>
<td>beta(1,3)</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>beta(1,4)</td>
<td>109.2</td>
<td>109.5</td>
<td>109.9</td>
<td>110.2</td>
<td>110.3</td>
<td>110.4</td>
</tr>
<tr>
<td>beta(1,5)</td>
<td>120.0</td>
<td>120.5</td>
<td>121.2</td>
<td>121.8</td>
<td>122.0</td>
<td>122.2</td>
</tr>
<tr>
<td>beta(2,5)</td>
<td>83.5</td>
<td>83.5</td>
<td>83.4</td>
<td>83.3</td>
<td>83.2</td>
<td>83.2</td>
</tr>
<tr>
<td>beta(5,1)</td>
<td>209.3</td>
<td>201.6</td>
<td>189.2</td>
<td>178.8</td>
<td>174.0</td>
<td>170.8</td>
</tr>
</tbody>
</table>
§5. Applications

In this section, we will first illustrate the effectiveness of the CUSUM chart using 2 simulated data sets and then analyze the performances of 7 surgeons based on a real data set.

For the simulated data sets, we assume that the distribution of $S/100$ follows a beta$(1, 3)$ distribution and the control limits is set such that the in-control ARL is 100. The first simulated data set consists of 100 patients. The first 50 patients are assumed to be operated on by a surgeon with a performance characterized by $R_0^* = R_1^* = 1$. The last 50 patients are assumed to be operated on by the surgeon with a performance characterized by $R_0^* = R_1^* = 0.5$. This means there is a degradation in the performance of the surgeon for the second half of the patients. Figure 5.1 displays the CUSUM charts for detecting both degradation and improvement. The control limits are chosen such that the in-control ARL for each chart is 100. The figures show that there is no unusual activity for the first half of the data. For the second half of the data, the CUSUM chart for detecting degradation shows an upward trend and eventually signals at the 63th patient.

The second data set also consists of 100 patients. The first 50 patients are assumed to be operated on by a surgeon with a performance characterized by $R_0 = R_1 = 1$. The last 50 patients are assumed to be operated on by a surgeon with a performance characterized by $R_0 = R_1 = 2$. This means there is an improvement in the performance of the surgeon for the second half of the patients. Figure 5.2 displays the CUSUM charts for detecting both degradation and improvement. The control limit is chosen such that the in-control ARL for each chart is 100. The figure shows that there is no unusual activity for the first half of the data. For the second half of the data, the CUSUM chart for detecting degradation becomes much quieter. In contrast, the CUSUM chart for detecting degradation shows an upward trend and eventually signal at the 79th patient.
Figure 5.1. Plots of the CUSUM charts for detecting degradation and improvement when the risk distribution of $S/100$ is assumed to follow a beta($a = 1, b = 3$) distribution. The first 50 patients are assumed to be operated on by a surgeon with a performance characterized by $R_0 = R_1 = 1$. The last 50 patients are assumed to be operated on by the surgeon with a performance characterized by $R_0 = R_1 = 0.5$ which represents a degradation. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 5.2. Plots of CUSUM charts for detecting degradation and improvement when the risk distribution of $S/100$ is assumed to follow a beta($a = 1, b = 3$) distribution. The first 50 patients are assumed to be operated on by a surgeon with a performance characterized by $R_0 = R_1 = 1$. The last 50 patients are assumed to be operated on by the surgeon with a performance characterized by $R_0 = R_1 = 2.0$ which represents an improvement. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Next we will examine the performances of 7 surgeons based on a real data set using the CUSUM charting procedure. The data set consists of 6449 patients who underwent cardiac surgeries. The outcomes of the surgeries are grouped into 3 categories: full recovery, partial recovery and death. A surgeon who implements such a chart should be able to classify the 3 outcomes appropriately. Our classification of the 3 outcomes is an approximation since we do not have detailed information of the patients. A patient who died within 30 days is considered a death with $Y = 2$. A patient who died after 30 days when the data was collected is considered a partial recovery with $Y = 1$. A patient who survived when the data was collected is considered a full recovery with $Y = 0$. The risk distributions of the patients operated on by the 7 surgeons are shown in Figures 5.3–5.9. In general, the risk distributions are highly right skewed with the density decreasing with increasing risk. The beta density estimated using the method-of-moments estimators is also shown on each plot. The beta densities provide reasonable fit to the risk distributions. The average performance of the 7 surgeons is characterized by the proportional logistic regression model as described by equation (3.2) fitted using the entire data set. Our objective is to examine the historical performances of the 7 surgeons. For each of the 7 surgeons, both CUSUM charts for detecting degradation ($R = 0.5$) and improvement ($R = 2$) are plotted. In order to compare all the charts fairly, the chart limit for each chart is set at 3.5 times the standard deviation of the steady-state CUSUM statistic without a control limit. We set the limits to be relatively large in order to study the historical performances of the 7 surgeons. These chart limits may need to be set much smaller for online monitoring.

Figures 5.10–5.16 show the CUSUM charts for the 7 surgeons. Surgeon 1 operated on 1654 patients. Figure 5.10 shows that his performance is in control for approximating the first 200 patients. His performance then degraded for the next 900 patients but became in-control again for the rest of the patients. Surgeon 2 operated on 568 patients. Figure 5.11 shows that his performance is in control for
approximately the first half of the patients. His performance then degraded for the second half of the patients. Surgeon 3 operated on 986 patients. Figure 5.12 shows that his performance is in control for all the patients and shows an improvement from patient number 680 onwards. Surgeon 4 operated on 209 patients. Figure 5.13 shows that his performance is in control and stable throughout. Surgeon 5 operated on 781 patients. Figure 5.14 shows that his performance is in control for all the patients and shows improvements throughout. Surgeon 6 operated on 1645 patients. Figure 5.15 shows that his performance is in control for most of the patients and shows a significant improvement from patient number 1150 onwards. Surgeon 7 operated on 606 patients. Figure 5.16 shows that his performance in control for the most part but shows signs of degradation in the middle. In summary, Surgeons 1 and 5 show the worst and best historical performances respectively. For a future patient, it is the current performance of a surgeon that counts. As for their current performances, (i) Surgeon 2 is performing worse than average; (ii) Surgeons 3, 5 and 6 are performing above average with Surgeon 6 showing the best performance; (iii) Surgeons 1, 4 and 7 are performing close to the average. The CUSUM charting procedures show clearly the difference among the 7 surgeons and our analyses show that the procedures are useful in understanding differences in performances among the 7 surgeons.
Figure 5.3. Histograms of $S/69$ of patients operated on by Surgeon 1 with the best fitted beta density based on the method-of-moments estimators.

Figure 5.4. Histograms of $S/65$ of patients operated on by Surgeon 2 with the best fitted beta density based on the method-of-moments estimators.
Figure 5.5. Histograms of $S/67$ of patients operated on by Surgeon 3 with the best fitted beta density based on the method-of-moments estimators.

Figure 5.6. Histograms of $S/43$ of patients operated on by Surgeon 4 with the best fitted beta density based on the method-of-moments estimators.
Figure 5.7. Histograms of $S/47$ of patients operated on by Surgeon 5 with the best fitted beta density based on the method-of-moments estimators.

Figure 5.8. Histograms of $S/59$ of patients operated on by Surgeon 6 with the best fitted beta density based on the method-of-moments estimators.
Figure 5.9. Histograms of $S/71$ of patients operated on by Surgeon 7 with the best fitted beta density based on the method-of-moments estimators.
Figure 5.10. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 1. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 5.11. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 2. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 5.12. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 3. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 5.13. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 4. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 5.14. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 5. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
(a) CUSUM chart for detecting degradation ($R = 0.5$)

(b) CUSUM chart for detecting improvement ($R = 2$)

(c) CUSUM chart for detecting degradation ($R = 0.5$)

(d) CUSUM chart for detecting improvement ($R = 2$)

Figure 5.15. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 6. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 5.16. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 7. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
§6. Comparison with CUSUM Chart based on Binary Outcomes

In this section, we will compare our approach with that of Steiner et al. (2000) in which there are 2 outcomes: survival \((Y_b = 0)\) and death \((Y_b = 1)\). This binary case is a special case of our approach with \(J = 1\). Figures 6.1–6.7 display the CUSUM charts of the 7 surgeons based on 2 outcomes. In order to compare all the charts fairly, the chart limit for each chart is also set at 3.5 times the standard deviation of the steady-state CUSUM statistic.

For detecting degradation, Figure 6.8 contains the plots of \(W(y, s)\) against the Parsonnet score \(s\) for (i) \(y = 0, 1, 2\) when \(J = 2\) and \(R_0 = R_1 = 0.5\) (solid lines), (ii) \(y_b = 0, 1\) when \(J = 1\) and \(R_0 = 0.5\) (dashed lines). The lines for \(y_b = 1\) and \(y = 2\) are very similar which means that the penalty for a patient who died is very similar for both procedures. In fact the line for \(y_b = 1\) is almost uniformly above the line for \(y = 2\). The outcomes \(y = 0\) and \(y = 1\) for the 3-outcome case collapse into the outcome \(y_b = 0\) for the binary case. For a patient who survives an operation, there is a bigger difference in the reward (see lines for \(y_b = 0\) and \(y = 0\)) with the reward for the binary case being uniformly smaller than that of the 3-outcome case (the line for \(y_b = 0\) is uniformly above the line for \(y = 0\)). The most significant difference however is that there is a very big drop from the line \(y = 1\) to the line \(y_b = 0\).

For Surgeon 1, a comparison of Figures 6.1 (a) and 5.10 (a) reveals that the plot in 6.1 (a) is substantially lower than that in 5.10 (a). This can be explained with what we have observed in Figure 6.8. The overall percentage of outcome \(Y = 1\) is 5.6%. The percentage of outcome \(Y = 1\) is 6.5% from patient number 401 to 800, and 2.5% from patient number 801 to 1000. The higher percentage of 6.5% causes a much bigger drop in the plot from patient number 401 to 800 than that from patient number 801 to 1000. This resulted in reducing the 6 signals for the 3-outcome case to 3 for the binary case as shown in Figures 5.10 (c) and 6.1 (c). Thus, the chart for the binary case is less sensitive in detecting degradation in the presence of substantial proportion of partial recoveries.
For Surgeon 2, the effect of outcome \( Y = 1 \) is clearly revealed in the latter part of the plots as shown in Figures 6.2 (a) and 5.11 (a). The overall percentage of outcome \( Y = 1 \) is 8.4%. The percentage of outcome \( Y = 1 \) is 12.0% from patient number 250 to 400 and this causes a big drop in the plot. This also resulted in reducing the 3 signals for the 3-outcome case to 1 for the binary case as shown in Figures 5.11 (c) and 6.2 (c). Again, the chart for the binary case is less sensitive in detecting degradation in the presence of substantial proportion of partial recoveries. Furthermore, the effect is made more pronounced by the fact that the distribution of the Parsonnet score is highly right skewed.

For detecting improvement, Figure 6.9 shows the plots of \( W(y, s) \) against Parsonnet score \( s \) for (i) \( y = 0, 1, 2 \) when \( J = 2 \) and \( R_0 = R_1 = 2 \) (solid lines), (ii) \( y_b = 0, 1 \) when \( J = 1 \) and \( R_0 = 2 \) (dashed lines). The lines for \( y_b = 1 \) and \( y = 2 \) are very similar which means that the penalty for a patient who died is very similar. In fact the line for \( y_b = 1 \) is almost uniformly below the line for \( y = 2 \). The outcomes \( y = 0 \) and \( y = 1 \) for the 3-outcome case collapse into one outcome \( y_b = 0 \) for the binary case. For a patient who survives an operation, there is a bigger difference in the reward (see lines for \( y_b = 0 \) and \( y = 0 \)) with the reward for the binary case being smaller compared to the 3-outcome case (the line for \( y_b = 0 \) is uniformly below the line for \( y = 0 \)). The most significant difference however is that there is a very big rise from the line \( y = 1 \) to the line \( y_b = 0 \).

For Surgeon 6, it can be seen from Figures 6.6 (b) and 5.15 (b) that a signal was issued by the CUSUM chart for the binary case at the 638th patient but this was not issued by the CUSUM chart for the 3-outcome case. This can also be explained with what we have observed in Figure 6.9. The overall percentage of outcome \( Y = 1 \) is 3.0%. The percentage of outcome \( Y = 1 \) is 7.0% from patient number 550 to 650. The substantial proportion of partial recoveries causes the CUSUM chart for the binary case to signal. This signal is a result of loss of information by treating a partial recovery as a full recovery. Furthermore, the effect is made more pronounced
by the fact that the distribution of the Parsonnet score is highly right skewed.

For Surgeon 3, it can be seen from Figures 6.3 (b) and 5.12 (b) that the CUSUM chart for the binary case did not issue any signal whereas the CUSUM chart for the 3-outcome case issues 2 signals at patient number 760 and 980. This is contrary to what we observed for Surgeon 6. The overall percentage of $Y = 1$ is 4.1% and the percentage of $Y = 1$ from 700 to 986 patient is only 2.1%. The overall percentage of $Y = 0$ is 90.6% and the percentage of $Y = 0$ from 700 to 986 patient is 94.1%. Figure 6.9 shows that converting $Y = 1$ to $Y_b = 0$ will cause the CUSUM statistic to rise in the CUSUM chart for the binary case. However, the line $Y = 0$ is uniformly greater than $Y_b = 0$, this causes the CUSUM statistic to drop in the CUSUM chart for the binary case. In the presence of a larger proportion (94.1%) of $Y = 0$ and a smaller proportion (2.1%) of $Y = 1$, the net result is a drop of CUSUM statistic for the binary case. The same explanation is true for Surgeons 5 and 6.
Figure 6.1. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 1 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.

Figure 6.2. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 2 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Chart (c) and (d) reset the CUSUM statistic to 0 after a signal is issued. charts (c) and (d) are useful for surgeon 3 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued. Plots of the CUSUM charts for detecting both degradation and improvement. 

**Figure 6.3.** Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 3 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 6.4. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 4 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 6.5. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 5 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 6.6. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 6 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.

Figure 6.7. Plots of the CUSUM charts for detecting both degradation and improvement for surgeon 7 when the outcome of the surgery is assumed to be binary. Charts (c) and (d) reset the CUSUM statistic to 0 after a signal is issued.
Figure 6.8. Plots of $W(y, s)$ against the Parsonnet score $s$ for (i) $y = 0, 1, 2$ when $J = 2$ and $R_0 = R_1 = 0.5$ (solid lines), (ii) $y_b = 0, 1$ when $J = 1$ and $R_0 = 0.5$ (dashed lines).

Figure 6.9. Plots of $W(y, s)$ against the Parsonnet score $s$ for (i) $y = 0, 1, 2$ when $J = 2$ and $R_0 = R_1 = 2$ (solid lines), (ii) $y_b = 0, 1$ when $J = 1$ and $R_0 = 2$ (dashed lines).
§7. Conclusion

Steiner et al. (2000) first developed a risk-adjusted CUSUM charting procedure for monitoring binary outcomes: survival and dead. Using the preoperative risk of a patient, the binary outcome is converted into a continuous penalty-reward score. It is naive to regard a patient who has a full recovery the same as another patient who survived but remained bed-ridden for life. For a patient who survives an operation, there can be many different grades of recovery. It thus makes sense to consider a risk-adjusted CUSUM charting procedure based on more than 2 outcomes. In this chapter, we have developed a risk-adjusted CUSUM charting procedure for monitoring the performance of a surgical process with 3 or more outcomes. Based on a historical data set, a proportional odds logistic regression model is used to estimate the probabilities of the outcomes. The properties of the penalty-reward score are then investigated and conditions are found to ensure that there is a proper ordering for the penalty-reward score according to different severities of outcomes. We have also adapted a procedure based on the collocation method (Knoth, 2005) for approximating the ARL. Sensitivity analyses show that as long as the risk distribution is close to the true distribution, the change in the ARL is small. If there is a big departure from the true distribution, the effect on the ARL can be significant. Based on an analysis of a real data set consisting of 7 surgeons using our risk-adjusted CUSUM procedure, we are able to show clearly differences in their performances. Our procedure is thus a valuable tool for analyzing and comparing the performances of surgeons. Treating a partial recovery as a full recovery results in a big drop in the penalty score and a big rise in the reward score. By using 2 outcomes instead of 3 outcomes, we found that the CUSUM chart for the binary case becomes less sensitive in detecting degradation. In addition, the CUSUM chart for the binary case generally issues more false alarms. Finally, it should be noted that the underlying risk distribution of the patients has an effect on the CUSUM chart. Hence, it is important and necessary to also monitor the risk distribution of
the patients simultaneously to prevent any false alarm due to changes in the risk distribution. This will be investigated in future research.
Appendix A: CDF and PDF of $W(Y,S)$

Let \( g_k(s) = \log[\pi_k^A(s)/\pi_k(s)] \), \( 1(\cdot) \) be the indicator function.

If \( R_0 = R_1 = \cdots = R_{J-1} = R < 1 \). Then the cdf \( F_W(\cdot) \) and the pdf \( f_W(\cdot) \) of \( W(Y,S) \) are given by

\[
F_W(w) = \sum_{k=0}^{J} \left\{ \int_{s_k(w)}^{100} \pi_k^*(s)f(s)ds \cdot 1(l_k < w < u_k) + \int_{0}^{100} \pi_k^*(s)f(s)ds \cdot 1(w \geq u_k) \right\},
\]

and

\[
f_W(w) = -\sum_{k=0}^{J} \pi_k^*(s_k(w))f(s_k(w))s'_k(w)1(l_k < w < u_k),
\]

where \( l_k = g_k(100) = \min_{s \in [0,100]} g_k(s) \), \( u_k = g_k(0) = \max_{s \in [0,100]} g_k(s) \), and \( s_k(w) \in [0,100] \) such that \( g_k(s_k(w)) = w \).

If \( R_0 = R_1 = \cdots = R_{J-1} = R > 1 \). Then the cdf \( F_W(\cdot) \) and the pdf \( f_W(\cdot) \) of \( W(Y,S) \) are given by

\[
F_W(w) = \sum_{k=0}^{J} \left\{ \int_{0}^{s_k(w)} \pi_k^*(s)f(s)ds \cdot 1(l_k < w < u_k) + \int_{0}^{100} \pi_k^*(s)f(s)ds \cdot 1(w \geq u_k) \right\},
\]

and

\[
f_W(w) = \sum_{k=0}^{J} \pi_k^*(s_k(w))f(s_k(w))s'_k(w)1(l_k < w < u_k),
\]

where \( l_k = g_k(0) = \min_{s \in [0,100]} g_k(s) \), \( u_k = g_k(100) = \max_{s \in [0,100]} g_k(s) \), and \( s_k(w) \in [0,100] \) such that \( g_k(s_k(w)) = w \).

Proof. Without loss of generality, we only need to prove the first part.

If \( R_0 = R_1 = \cdots = R_{J-1} = R < 1 \), then \( W(Y,S) \) is a decreasing function of \( S \) conditional on \( Y \). Then, \( g_k(s), k = 0, \cdots, J \) are all decreasing functions. Thus, we have

\[
l_k = g_k(100) = \min_{s \in [0,100]} g_k(s),
\]

\[
u_k = g_k(0) = \max_{s \in [0,100]} g_k(s).
\]
Let $s_k(w) \in [0, 100]$ such that $g_k(s_k(w)) = w$, then it is the inverse function of $g_k(s)$. Therefore,

$$F_W(w) = P(W(Y, S) \leq w) = \sum_{k=0}^{J} P(W(Y, S) \leq w, Y = k)$$

$$= \sum_{k=0}^{J} \int_{g_k(s) \leq w} \pi_k^*(s) f(s) ds$$

$$= \sum_{k=0}^{J} \left\{ \int_{s_k(w)}^{100} \pi_k^*(s) f(s) ds \cdot 1(l_k < w < u_k) + \int_{0}^{100} \pi_k^*(s) f(s) ds \cdot 1(w \geq u_k) \right\}. $$

Taking the first derivative of the cdf of $W(Y, S)$, we obtain the pdf

$$f_W(w) = -\sum_{k=0}^{J} \pi_k^*(s_k(w)) f(s_k(w)) s_k'(w) 1(l_k < w < u_k).$$
Appendix B: Collocation Method

The collocation method is an accurate method of approximating the ARL of a charting procedure by numerically solving the integral equation introduced by Page (1954). Knoth (2005) demonstrated that the collocation method is fast and accurate in assessing the ARL performance of an exponentially weighted moving average (EWMA) control chart for monitoring the variance of normally distributed data. Consider a CUSUM chart obtained by plotting

\[ C_n = \max(0, C_{n-1} + W_n), \]

against the sample number \( n \) where \( C_0 = u, 0 \leq u < h \) and \( W_n \) is the log likelihood ratio statistic

\[ W_n = \log \left\{ \frac{f(X_n; \theta_A)}{f(X_n; \theta_0)} \right\}. \]

Let \( L(u) \) denote the ARL of the CUSUM chart that starts at \( C_0 = u \), then the integral equation for the ARL derived by Page (1954) is given as

\[ L(u) = 1 + L(0)F_W(-u) + \int_0^h L(x)f_W(x-u)dx, \]

where \( F_W(\cdot) \) and \( f_W(\cdot) \) are the cdf and pdf of \( W(Y, S) \) respectively.

The collocation method is to approximate \( L(u) \) by \( \sum_{j=1}^{N} c_j T_j(u) \) for \( u \in [0, h] \), where \( T(\cdot) \) is a set of \( N \) independent interpolating functions, and \( c_j \)'s are unknown constant. To solve \( c_j \)'s, we choose a set of \( N \) nodes in the domain \([0, h]\), and then solve the resulting system of linear equations as discussed in Hackbusch (1995). According to Knoth (2005), the Chebychev polynomials \( T_j(z) = \cos((j-1) \arccos(z)) \), \( z \in [-1, 1], j = 1, \ldots, N \) provide stable numerical quadratures. The Chebychev nodes are given as \( z_i = \cos((2i-1)\pi/2N), i = 1, \ldots, N \). The Chebychev polynomials and nodes are designed for the domain \([-1, 1]\). In order to use them in the domain \([0, h]\), we consider the Chebychev polynomials and nodes as \( T_j(z) = \cos((j-1) \arccos(\frac{2z-h}{h})) \), \( z \in [0, h], j = 1, \ldots, N \) and \( z_i = \frac{h}{2}[1 + \cos((2i-
1) $\pi/2N)$, $i = 1, \cdots, N$. Then, $c_j$’s can be obtained by solving the following system of linear equations:

$$
\sum_{j=1}^{N} c_j T_j(z_i) = 1 + F_W(-z_i) \sum_{j=1}^{N} c_j T_j(0) + \sum_{j=1}^{N} c_j \int_{0}^{h} T_j(x) f_W(x - z_i) dx,
$$

(B1)

for $i = 1, \cdots, N$. The integral on the right-hand side can be approximated using the Gauss-Legendre quadratures (Abramowitz and Stegun, 1968).

Note that the pdf of $W(Y, S)$ is

$$
f_W(w) = -\sum_{k=0}^{J} \pi_k^*(s_k(w)) f(s_k(w)) s_k'(w) 1(l_k < w < u_k).
$$

The function as a whole is not smooth enough for the Gauss-Legendre quadratures to estimate accurately. However, the component $\pi_k^*(s_k(w)) f(s_k(w)) s_k'(w) 1(l_k < w < u_k)$ is smooth enough for the Gauss-Legendre quadratures to estimate accurately. Hence, the integral on the right-hand of (B1) is approximated by

$$
\int_{0}^{h} T_j(x) f_W(x - z_i) dx = -\sum_{k=0}^{J} \int_{0}^{h} T_j(x) \pi_k^*(s_k(w)) f(s_k(w)) s_k'(w) 1(l_k < w < u_k) dx,
$$

and each integral in the right-hand can be approximated by the Gauss-Legendre quadratures.
Appendix C: R Codes for Binary Outcome Case

1 library(MASS);
2 sgh=read.table("SGH1_data.txt");
3 record=sgh[,13];parsonnet=sgh[,15];
4 nrecord=length(record);
5 newrecord=rep(0,nrecord);
6 for(i in 1:nrecord)
7 {if(record[i]<31&record[i]>-1) newrecord[i]=1
8 else newrecord[i]=0
9 }
10 response=as.factor(newrecord);
11 modelt=glm(response~parsonnet,family=binomial("logit"));
12 alpha=as.numeric(-modelt$coef[1]);
13 beta=as.numeric(-modelt$coef[2]);
14
15 ###########################distribution of parsonnet scores###########################
16 f_p=function(par){
17 den_y=dbeta(par/100,1,3)/100;
18 return(den_y);
19 }
20 ################################################################
21 getprob=function(par,r0){
22 p1=predict(modelt,data.frame(parsonnet=par),"response");
23 p0=1-p1;
24 prob_n=c(p0,p1);
25 as.vector(prob_n)->prob_n;
26
27 p_n_0=prob_n[1];
28 p_n_1=prob_n[2];
29
30 p_a_0=r0*p_n_0/(1-p_n_0+r0*p_n_0);
31 p_a_1=1-p_a_0;
32
33 prob_a=c(p_a_0,p_a_1);
34 return(prob_a);
35 }


getpi=function(par,out,r0){
  prob_n=getprob(par,r0);
  result=prob_n[out+1];
  return(result);
}

getw=function(par,out,r_a){
  prob_n=getprob(par,1);
  prob_a=getprob(par,r_a);
  inc=log(prob_a[out+1]/prob_n[out+1]);
  return(inc);
}

#Inverse function of g#
getw_I=function(w,out,r_a){
  sp1=0;
  sp2=100;
  w1=getw(sp1,out,r_a);
  w2=getw(sp2,out,r_a);
  w_min=min(w1,w2);
  w_max=max(w1,w2);

  if(r_a<1) {
    if(w>=w_min&w<=w_max){
      spt=(sp1+sp2)/2;
      wt=getw(spt,out,r_a);
      while(abs(sp2-sp1)>0.0000000001){
        if(wt<w) sp2=spt else sp1=spt
        spt=(sp1+sp2)/2;
        wt=getw(spt,out,r_a);
      }
      return(spt);
    }
  } else return(NULL)
}
Risk-Adjusted CUSUM Charts and SMR

Tang Xu

```r
72 } 
73 
74 else { 
75 if(w>=w_min&w<=w_max){ 
76 spt=(sp1+sp2)/2; 
77 wt=getw(spt,out,r_a); 
78 
79 while(abs(sp2-sp1)>0.0000000001){ 
80 if(wt<w) sp1=spt else sp2=spt 
81 
82 spt=(sp1+sp2)/2 
83 wt=getw(spt,out,r_a) 
84 } 
85 
86 return(spt); 
87 } 
88 
89 else return(NULL) 
90 } 
91 } 
92 
93 #############Gaussian quatrature for CDF OF W################## 
94 integ_c=function(out,int_l,int_u,r0,N_Gaussian){ 
95 if(N_Gaussian==40){ 
96 p=numeric(20); 
97 p[1]=0.9982377097105592; p[2]=0.9907262386994570; 
98 p[3]=0.9772599499837742; p[4]=0.9579168192137916; 
99 p[5]=0.9328128082786765; p[6]=0.9020988069688742; 
100 p[7]=0.8659595032122595; p[8]=0.8246122308333116; 
101 p[9]=0.7783056514265194; p[10]=0.7273182551899271; 
102 p[11]=0.6719566846141795; p[12]=0.6125538896679802; 
103 p[13]=0.5494671250951282; p[14]=0.4830758016861787; 
104 p[15]=0.4137792043716050; p[16]=0.3419940908257584; 
105 p[17]=0.2681521850072536; p[18]=0.1926975807013710; 
106 p[19]=0.1160840706752552; p[20]=0.0387724175060508; 
107 p=c(-p,rev(p)); 
63
```
w=numeric(20);
w[1]=0.004521277098533191;  w[2]=0.01049828491416695;
w[3]=0.0164210583190788;  w[4]=0.02224584918567327;
w[5]=0.02793706763507223;  w[6]=0.03346019528254784;
w[7]=0.03878216797447201;  w[8]=0.04387090819867327;
w[9]=0.04895607663507223;  w[10]=0.05322784698939362;
w[11]=0.05743976909939155;  w[12]=0.06130624249292893;
w[13]=0.06480401345660103;  w[14]=0.06791204581523390;
w[15]=0.07061164739182877;  w[16]=0.07288658239580405;
w[17]=0.07472316905796826;  w[18]=0.07611036190062624;
w[19]=0.077039881816424796; w[20]=0.07750594797842481;
w=c(w,rev(w));
}

if(N_Gaussian==80){
p=numeric(40);
p[1]=0.01951138325679397654;  p[2]=0.058504437152420668629;
p[3]=0.097408398441584599063;  p[4]=0.136164022809143886599;
p[5]=0.174712291832646812559;  p[6]=0.212994502857666132572;
p[7]=0.25095235892272120493;  p[8]=0.288528054884511853109;
p[9]=0.325664370747701914619;  p[10]=0.36230475349487315619;
p[11]=0.398393405881969227024;  p[12]=0.433875370831756093062;
p[13]=0.46869661517054477036;  p[14]=0.5028041118888888798594;
p[15]=0.53614592089713193202; p[16]=0.568671268122709784725;
p[17]=0.600330622829751743155;  p[18]=0.631075773046871966248;
p[19]=0.6608598986119801736;  p[20]=0.6896376443422027600771;
p[21]=0.71736518536209988254;  p[22]=0.744000297583597272317;
p[23]=0.769502420135041373866;  p[24]=0.79383271750460549949;
p[25]=0.816954138681463470371;  p[26]=0.838831473580255275617;
p[27]=0.85943140666311096977;  p[28]=0.87872256767821382704;
p[29]=0.89667579438770683194;  p[30]=0.913263102571757654165;
p[31]=0.92845977172445795953;  p[32]=0.94224276130987267452;
p[33]=0.954590766343634905493;  p[34]=0.965485089043799251452;
p[35]=0.974909140585727793386;  p[36]=0.982849572738629070418;
p[37]=0.98929130249975531027;  p[38]=0.994227540965688277829;
p[39]=0.997649864398237688890;  p[40]=0.999553822651630629880;
\begin{verbatim}
144  p=c(-rev(p),p);
145
146  w=numeric(40);
147  w[1]=0.039017813656306654811;  w[2]=0.038958395962769531199;
148  w[3]=0.038839651059051968932;  w[4]=0.038661759774076463327;
149  w[5]=0.038424993006959423185;  w[6]=0.038129711314477638344;
150  w[7]=0.037776364362001397490;  w[8]=0.037365490238730490027;
151  w[9]=0.036897714638276008839;  w[10]=0.036373749905835978044;
152  w[11]=0.035794393953416054603; w[12]=0.035160529044747593496;
153  w[13]=0.034473120451753928794; w[14]=0.033733214984611522817;
154  w[15]=0.032941939397645401383; w[16]=0.03210049867348773148;
155  w[17]=0.031210174188114701642; w[18]=0.030273221759557908661;
156  w[19]=0.029288369583267847693; w[20]=0.02825981605726862397;
157  w[21]=0.027188227500486380674; w[22]=0.026075235767565117903;
158  w[23]=0.024922535764115491105; w[24]=0.023731882865930101293;
159  w[25]=0.022505090246332461926; w[26]=0.021244026115782006389;
160  w[27]=0.019950610878141998929; w[28]=0.0186268142082899031429;
161  w[29]=0.017274652056269306359; w[30]=0.015896183583725668042;
162  w[31]=0.014493508040509076117; w[32]=0.013068761592401339294;
163  w[33]=0.01162411420797826916; w[34]=0.010161766041103064521;
164  w[35]=0.008683945269260858426; w[36]=0.007192904768117312753;
165  w[37]=0.005690922451403198649; w[38]=0.004180313124694895237;
166  w[39]=0.002663533589512681669; w[40]=0.001144950003186941534;
167  w=c(rev(w),w);
168 }
169
170 if(int_l>=int_u) int=0
171 else{
172  p_n=(int_u-int_l)/2*(p-1)+int_u;
173  fp=numeric(N_Gaussian);
174  for(i in 1:N_Gaussian) fp[i]=getpi(p_n[i],out,r0)*f_p(p_n[i]);
175  int=(int_u-int_l)/2*(w%*%fp);
176 }
177 int=as.numeric(int);
178 return(int)
179 }
\end{verbatim}
F_W_t=function(w,out,r0,r_a,N_Gaussian){
  p_t_1=0;
  p_t_2=100;
  w_t_1=getw(p_t_1,out,r_a);
  w_t_2=getw(p_t_2,out,r_a);
  w_l=min(w_t_1,w_t_2);
  w_u=max(w_t_1,w_t_2);
  if(r_a<1){
    if(w<w_l) return(0)
    else{ if(w>w_u) return(integ_c(out,p_t_1,p_t_2,r0,N_Gaussian))
      else{ p_t_c=getw_I(w,out,r_a)
        return(integ_c(out,p_t_c,p_t_2,r0,N_Gaussian))
      }
    }
  } else {
    if(w<w_l) return(0)
    else{ if(w>w_u) return(integ_c(out,p_t_1,p_t_2,r0,N_Gaussian))
      else{ p_t_c=getw_I(w,out,r_a)
        return(integ_c(out,p_t_1,p_t_c,r0,N_Gaussian))
      }
    }
  }
}

F_W=function(w,r0,r_a,N_Gaussian){
  J=2;
  temp=rep(0,J);
  for(i in 1:J) temp[i]=F_W_t(w,i-1,r0,r_a,N_Gaussian);
  FW=sum(temp);
  return(FW);
}
\text{PDF OF W}\n
\begin{verbatim}
dw=function(par,out,r_a){
  prob_n=getprob(par,1);
  prob_a=getprob(par,r_a);
  temp1_n=sum(prob_n[0:(out+1)]);
  temp2_n=sum(prob_n[0:out]);
  temp1_a=sum(prob_a[0:(out+1)]);
  temp2_a=sum(prob_a[0:out]);
  result=beta*(temp1_n+temp2_n-temp1_a-temp2_a);
  return(result);
}

f_W_t=function(w,out,r0,r_a){
  p_t_1=0;
  p_t_2=100;
  w_t_1=getw(p_t_1,out,r_a);
  w_t_2=getw(p_t_2,out,r_a);
  w_l=min(w_t_1,w_t_2);
  w_u=max(w_t_1,w_t_2);
  if(r_a<1){
    if(w<=w_l|w>=w_u) return(0)
    else{ p_t_c=getw_I(w,out,r_a)
      result=-getpi(p_t_c,out,r0)*f_p(p_t_c)/dw(p_t_c,out,r_a);
      return(result);
    }
  }
  else {
    if(w<=w_l|w>=w_u) return(0)
    else{ p_t_c=getw_I(w,out,r_a)
      result=-getpi(p_t_c,out,r0)*f_p(p_t_c)/dw(p_t_c,out,r_a);
      return(result);
    }
  }
}
\end{verbatim}
result=getpi(p_t_c,out,r0)*f_p(p_t_c)/dw(p_t_c,out,r_a);
return(result);
}
}

f_W=function(w,r0,r_a){
  J=2;
  temp=rep(0,J);
  for(i in 1:J) temp[i]=f_W_t(w,i-1,r0,r_a);
  fW=sum(temp);
  return(fW);
}

##############Chebyshev polynomials###################################
T=function(z,j,h){
t=cos((j-1)*acos((2*z-h)/h));
return(t)
}

#####nodes ############################################################
z=function(i,h,N){
  z=h/2*(1+cos((2*i-1)*pi/(2*N)));
  return(z)
}

nodes=function(h,N){
  z=numeric(N);
  for(i in 1:N) z[i]=z(i,h,N)
  return(z)
}

#############Gaussian quadrature#######################################
integ_t=function(out,h,j,loc,r0,r_a,N_Gaussian){
  if(N_Gaussian==40){

68
p = numeric(20);

p[1] = 0.9982377097105592; p[2] = 0.9907262386994570;
p[3] = 0.977259499837742; p[4] = 0.9579168192137916;
p[5] = 0.9328128082786765; p[6] = 0.9020988069688742;
p[7] = 0.8659595032122595; p[8] = 0.824612230833116;
p[9] = 0.7783056514265194; p[10] = 0.7273182551899271;
p[11] = 0.6719566846141795; p[12] = 0.6125538896679802;
p[13] = 0.5494671250951282; p[14] = 0.4830758016861787;
p[15] = 0.4137792043716050; p[16] = 0.341994090825784;
p[17] = 0.2681521850072536; p[18] = 0.1926975807013710;
p[19] = 0.1160840706752552; p[20] = 0.0387724175060508;
p = c(-p, rev(p));

w = numeric(20);
w[1] = 0.004521277098533191; w[2] = 0.01049828453115281;
w[3] = 0.0164210583190788; w[4] = 0.02224584919416695;
w[5] = 0.02793700698002340; w[6] = 0.03346019528254784;
w[7] = 0.03878216797447201; w[8] = 0.04387090818567327;
w[9] = 0.04869580763507223; w[10] = 0.05322784698393682;
w[11] = 0.05743976909939155; w[12] = 0.06130624249292893;
w[13] = 0.06480401345660103; w[14] = 0.06791204581523390;
w[15] = 0.07061164739128677; w[16] = 0.07288658239580405;
w[17] = 0.07472316905796826; w[18] = 0.07611036190062624;
w[19] = 0.07703981816424796; w[20] = 0.07750594797842481;
w = c(w, rev(w));

if (N_Gaussian == 80) {
  p = numeric(40);

  p[1] = 0.01951138325679397654; p[2] = 0.058504437152420668629;
p[3] = 0.097408398441584599063; p[4] = 0.136164022809143886559;
p[5] = 0.174712291832646812559; p[6] = 0.212994502857666132572;
p[7] = 0.250952358392272120493; p[8] = 0.288528054884511853109;
p[9] = 0.32566437074770194619; p[10] = 0.362304753499487315619;
p[11] = 0.398339405581969227024; p[12] = 0.433875370831756093062;
p[13] = 0.46869661517054477036; p[14] = 0.50280411188784987594;
p[15] = 0.536145920897131932020; p[16] = 0.568671268122709784725;
}
\[ p[17]=0.600330622829751743155; \]
\[ p[18]=0.63107577304650449949; \]
\[ p[19]=0.660859898986119801736; \]
\[ p[20]=0.689637644342027600771; \]
\[ p[21]=0.717365185362099880254; \]
\[ p[22]=0.744000297583597272317; \]
\[ p[23]=0.769502420135041373871; \]
\[ p[24]=0.816954138681463470371; \]
\[ p[25]=0.838831473580255275617; \]
\[ p[26]=0.85943140666311096977; \]
\[ p[27]=0.87822567678213828704; \]
\[ p[28]=0.913263102571757654165; \]
\[ p[29]=0.9284598771724475953; \]
\[ p[30]=0.94224276130872674752; \]
\[ p[31]=0.954590766343634905493; \]
\[ p[32]=0.965485089043799254152; \]
\[ p[33]=0.974909140585727793386; \]
\[ p[34]=0.9824957238629070418; \]
\[ p[35]=0.9892913024975531027; \]
\[ p[36]=0.99422754096588277829; \]
\[ p[37]=0.997649863498237688900; \]
\[ p[38]=0.999553822651630629880; \]
\[ p[c(-rev(p),p)]; \]
\[ w=numeric(40); \]
\[ w[1]=0.039017813656306654811; \]
\[ w[2]=0.038958395962769531199; \]
\[ w[3]=0.038839651059051968932; \]
\[ w[4]=0.038661759774076463272; \]
\[ w[5]=0.03842499300695423185; \]
\[ w[6]=0.0381297113144763844; \]
\[ w[7]=0.037776364362001397490; \]
\[ w[8]=0.03736549023873049027; \]
\[ w[9]=0.036897714638276008839; \]
\[ w[10]=0.03637349905835978044; \]
\[ w[11]=0.035794393953416054603; \]
\[ w[12]=0.035160529044747593496; \]
\[ w[13]=0.034473120451753928794; \]
\[ w[14]=0.033733214984611522817; \]
\[ w[15]=0.0329419397645401383; \]
\[ w[16]=0.032100498673487773148; \]
\[ w[17]=0.0312101748114701642; \]
\[ w[18]=0.03027321759557980661; \]
\[ w[19]=0.029283869583267847693; \]
\[ w[20]=0.028259816057276862397; \]
\[ w[21]=0.027188227500486380674; \]
\[ w[22]=0.026075235767565117903; \]
\[ w[23]=0.02492253576411591105; \]
\[ w[24]=0.02373182865930101293; \]
\[ w[25]=0.02250509024633241926; \]
\[ w[26]=0.02124402611578200639; \]
\[ w[27]=0.01950610878141998929; \]
\[ w[28]=0.018626814208299031429; \]
\[ w[29]=0.017274652056269306359; \]
\[ w[30]=0.015896183583725668042; \]
\[ w[31]=0.014493508040509076117; \]
\[ w[32]=0.013068761592401339294; \]
\[ w[33]=0.011624114120797826916; \]
\[ w[34]=0.01016176604110364521; \]
\[ w[35]=0.008683945269260585426; \]
\[ w[36]=0.007192904768117312753; \]
\[ w[37]=0.005690922451403198649; \]
\[ w[38]=0.00418031324694895237; \]
\[ w[39]=0.002663535389512681669; \]
\[ w[40]=0.0011449500003186941534; \]
\[ w=c(rev(w),w); \]
p_t_1=0;
p_t_2=100;
w_t_1=getw(p_t_1,out,r_a);
w_t_2=getw(p_t_2,out,r_a);
w_l=min(w_t_1,w_t_2);
w_u=max(w_t_1,w_t_2);

int_l=0;int_u=h;
ll=w_l;ul=w_u;
int_l_n=max(int_l,ll+loc);
int_u_n=min(int_u,ul+loc);

if(int_l_n>=int_u_n) int=0
else{
p_n=(int_u_n-int_l_n)/2*(p-1)+int_u_n;
fp=numeric(N_Gaussian);
for(i in 1:N_Gaussian)
    fp[i]=T(p_n[i],j,h)*f_W_t(p_n[i]-loc,out,r0,r_a);
int=(int_u_n-int_l_n)/2*(w%*%fp);
}
int=as.numeric(int);
return(int);

integ_use=function(h,j,loc,r0,r_a,N_Gaussian){
J=2;
temp=rep(0,3);
for(i in 1:J)
temp[i]=integ_t(i-1,h,j,loc,r0,r_a,N_Gaussian);
result=sum(temp);
return(result);
}

getmatrix=function(h,r0,r_a,N,N_Gaussian_1,N_Gaussian_2){
z = nodes(h, N);
m = matrix(0, N, N);

for (i in 1:N)
  for (j in 1:N)
    m[i, j] = T(0, j, h) * F_W(-z[i], r0, r_a, N, Gaussian_1) +
    integ_use(h, j, z[i], r0, r_a, N, Gaussian_2) - T(z[i], j, h)

return(m)

getc = function(h, r0, r_a, N, Gaussian_1, Gaussian_2){
de = getmatrix(h, r0, r_a, N, Gaussian_1, Gaussian_2);
b = rep(-1, N);
c = solve(de, b);
return(c)
}

getalr = function(startpoint, h, r0, r_a, N, Gaussian_1, Gaussian_2){
u = startpoint;
c = getc(h, r0, r_a, N, Gaussian_1, Gaussian_2);
p_t = rep(0, N);
for (j in 1:N) p_t[j] = T(u, j, h);
arl = c %*% p_t;
arl = as.numeric(arl);
return(arl)
}
Appendix D: R Codes for 3-outcome Case

```r
library(MASS);
sgh=read.table("SGH1_data.txt");
record=sgh[,13];parsonnet=sgh[,15];
nrecord=length(record);
newrecord=rep(0,nrecord);
for(i in 1:nrecord)
  {if(record[i]<31&record[i]>-1) newrecord[i]=2
   else {if(record[i]>30) newrecord[i]=1
         else newrecord[i]=0
   }
}
response=as.factor(newrecord);
modelt=polr(response~parsonnet);
beta=as.numeric(-modelt$coeff);
alpha0=as.numeric(modelt$zeta[1]);
alpha1=as.numeric(modelt$zeta[2]);

###############distribution of parsonnet scores###############
f_p=function(par){
den_y=dbeta(par/100,1,3)/100;
return(den_y);
}

getprob=function(par,r0,r1){
prob_n=predict(modelt,data.frame(parsonnet=par),type="p");
as.vector(prob_n)->prob_n
p_n_0=prob_n[1];
p_n_1=prob_n[2];
p_n_2=prob_n[3];
p_a_0=r0*p_n_0/(1-p_n_0+r0*p_n_0);
p_a_2=1/r1*p_n_2/(1-p_n_2+1/r1*p_n_2);
p_a_1=1-p_a_0-p_a_2;
prob_a=c(p_a_0,p_a_1,p_a_2);
return(prob_a);
}
```

73
getpi=function(par,out,r0,r1){
  prob_n=getprob(par,r0,r1);
  result=prob_n[out+1];
  return(result);
}

getw=function(par,out,r_a){
  prob_n=predict(modelf, data.frame(parsonnet=par), type="p");
  as.vector(prob_n)->prob_n
  prob_a=getprob(par,r_a,r_a);
  inc=log(prob_a[out+1]/prob_n[out+1]);
  return(inc);
}

getw_I=function(w,out,r_a){
  sp1=0;
  sp2=100;
  w1=getw(sp1,out,r_a);
  w2=getw(sp2,out,r_a);
  w_min=min(w1,w2);
  w_max=max(w1,w2);
  if(r_a<1) {
    if(w>=w_min&w<=w_max){
      spt=(sp1+sp2)/2;
      wt=getw(spt,out,r_a);
    }
  }
  while(abs(sp2-sp1)>0.0000000001){
    if(wt<w) sp2=spt else sp1=spt
    spt=(sp1+sp2)/2
    wt=getw(spt,out,r_a)
  }
return(spt);
}

else return(NULL)

else {

if(w>=w_min&w<=w_max){
spt=(sp1+sp2)/2;
wt=getw(spt,out,r_a);

while(abs(sp2-sp1)>0.0000000001){
if(wt<w) sp1=spt else sp2=spt
spt=(sp1+sp2)/2
wt=getw(spt,out,r_a)
}

return(spt);
}

else return(NULL)

}

Gaussian quadrature for CDF of W

integ_c=function(out,int_l,int_u,r0,r1,N_Gaussian){

if(N_Gaussian==40){
p=numeric(20);
p[1]=0.9982377097105592; p[2]=0.9907262386994570;
p[3]=0.9772599499837742; p[4]=0.9579168192137916;
Risk-Adjusted CUSUM Charts and SMR

Tang Xu

108  \( p[5]=0.9328128082786765; \)  \( p[6]=0.9020988069688742; \)
109  \( p[7]=0.8659595032122595; \)  \( p[8]=0.8246122308333116; \)
110  \( p[9]=0.7783056514265194; \)  \( p[10]=0.7273182551899271; \)
111  \( p[11]=0.6719566846141795; \)  \( p[12]=0.612538896679802; \)
112  \( p[13]=0.5494671250951282; \)  \( p[14]=0.4830758016861787; \)
113  \( p[15]=0.4137792043716050; \)  \( p[16]=0.3419940908257584; \)
114  \( p[17]=0.2681521850072536; \)  \( p[18]=0.1926975807013710; \)
115  \( p[19]=0.1160840706752552; \)  \( p[20]=0.0387724175060508; \)
116  \( p=c(-p,rev(p)); \)
117
118  \( w=numeric(20); \)
119  \( w[1]=0.004521277098533191; \)  \( w[2]=0.01049828453115281; \)
120  \( w[3]=0.01642105838190788; \)  \( w[4]=0.02224584919416695; \)
121  \( w[5]=0.02793700698002340; \)  \( w[6]=0.03346019528254784; \)
122  \( w[7]=0.03878216797447201; \)  \( w[8]=0.04387090818567327; \)
123  \( w[9]=0.04869580763507223; \)  \( w[10]=0.053227846989393682; \)
124  \( w[11]=0.05743976909939155; \)  \( w[12]=0.06130624249292893; \)
125  \( w[13]=0.06480401345660103; \)  \( w[14]=0.06791204581523390; \)
126  \( w[15]=0.07061164739128677; \)  \( w[16]=0.07288658239580405; \)
127  \( w[17]=0.07472316905796826; \)  \( w[18]=0.07611036190062624; \)
128  \( w[19]=0.07703981816424796; \)  \( w[20]=0.07750594797842481; \)
129  \( w=c(w,rev(w)); \)
130
131  \( p=numeric(40); \)
132  \( p[1]=0.01951138325679397654; \)  \( p[2]=0.05850443715242066829; \)
133  \( p[3]=0.097408398441584599063; \)  \( p[4]=0.13616402280914388659; \)
134  \( p[5]=0.174712291832646812559; \)  \( p[6]=0.212994502857666132572; \)
135  \( p[7]=0.250952358392272120493; \)  \( p[8]=0.288528054884511853109; \)
136  \( p[9]=0.325664370747701914619; \)  \( p[10]=0.362304753499487315619; \)
137  \( p[11]=0.398393405881969227024; \)  \( p[12]=0.433875370831756093062; \)
138  \( p[13]=0.468696615170544477036; \)  \( p[14]=0.502804111888784987594; \)
139  \( p[15]=0.536145920897131932020; \)  \( p[16]=0.568671268122709784725; \)
140  \( p[17]=0.600330622829751743155; \)  \( p[18]=0.631075773046871966248; \)
141  \( p[19]=0.660859898986119801736; \)  \( p[20]=0.689637644342027600771; \)
Risk-Adjusted CUSUM Charts and SMR

Tang Xu

p[21]=0.717365185362099880254; p[22]=0.744000297583597272317;
p[23]=0.769502420135041373866; p[24]=0.793832717504605449949;
p[25]=0.816954138681463470371; p[26]=0.838831473580255275617;
p[27]=0.859431406663111096977; p[28]=0.878722567678213828704;
p[29]=0.896675579438770683194; p[30]=0.913263102571757654165;
p[31]=0.928459877172445795953; p[32]=0.942242761309872674752;
p[33]=0.954590766343634905493; p[34]=0.965485089043799251452;
p[35]=0.974909140585727793386; p[36]=0.982849572738629070418;
p[37]=0.989291302499755531027; p[38]=0.994227540965688277829;
p[39]=0.997649864398237688900; p[40]=0.999553822651630629880;
p=c(-rev(p),p);

w=numeric(40);
w[1]=0.039017813656306654811; w[2]=0.038958395962769531199;
w[3]=0.038839651059051968932; w[4]=0.038661759774076463327;
w[5]=0.038424993006959423185; w[6]=0.038129711314477638344;
w[7]=0.037776364362001397490; w[8]=0.037365490238730490027;
w[9]=0.036897714638276008839; w[10]=0.036373749905835978044;
w[11]=0.035794393953416054603; w[12]=0.035160529044747593496;
w[13]=0.034473120451753928794; w[14]=0.033733214984611522817;
w[15]=0.032941939397645401383; w[16]=0.03210049867348773148;
w[17]=0.031210174188114701642; w[18]=0.030272321759557908061;
w[19]=0.02928836958326784793; w[20]=0.02825981605726862397;
w[21]=0.027188227500486380674; w[22]=0.026075235767565117903;
w[23]=0.024922535764115491105; w[24]=0.023731882665930101293;
w[25]=0.022505090246332461926; w[26]=0.021244026115782006389;
w[27]=0.019950610878141199829; w[28]=0.018626814208299031429;
w[29]=0.017274652065269306359; w[30]=0.015896183583725668042;
w[31]=0.014493508040509076117; w[32]=0.013068761592401339294;
w[33]=0.0116244114120797826916; w[34]=0.010161766041103064521;
w[35]=0.008683945269260858426; w[36]=0.007192904768117312753;
w[37]=0.005690922451403198649; w[38]=0.004180313124694895237;
w[39]=0.002663533589512681669; w[40]=0.001144950003186941534;
w=c(rev(w),w);

}
if(int_l>=int_u) int=0
else{
p_n=(int_u-int_l)/2*(p-1)+int_u;
fp=numeric(N_Gaussian);
for(i in 1:N_Gaussian)
    fp[i]=getpi(p_n[i],out,r0,r1)*f_p(p_n[i]);
int=(int_u-int_l)/2*(w%*%fp);
}
int=as.numeric(int);
return(int)
}

### CDF of w

F_W_t=function(w,out,r0,r1,r_a,N_Gaussian){
p_t_1=0;
p_t_2=100;
w_t_1=getw(p_t_1,out,r_a);
w_t_2=getw(p_t_2,out,r_a);
w_l=min(w_t_1,w_t_2);
w_u=max(w_t_1,w_t_2);
if(r_a<1){
    if(w<w_l) return(0)
    else{ if(w>w_u) return(integ_c(out,p_t_1,p_t_2,r0,r1,N_Gaussian))
            else{ p_t_c=getw_I(w,out,r_a)
                    return(integ_c(out,p_t_1,p_t_c,r0,r1,N_Gaussian))
                }
        }
}
else {
    if(w<w_l) return(0)
    else{ if(w>w_u) return(integ_c(out,p_t_1,p_t_2,r0,r1,N_Gaussian))
            else{ p_t_c=getw_I(w,out,r_a)
                    return(integ_c(out,p_t_c,p_t_2,r0,r1,N_Gaussian))
                }
        }
    }
F_W=function(w,r0,r1,r_a,N_Gaussian){
    J=3;
    temp=rep(0,J);
    for(i in 1:J) temp[i]=F_W_t(w,i-1,r0,r1,r_a,N_Gaussian);
    FW=sum(temp);
    return(FW);
}

#PDF OF W

dw=function(par,out,r_a){
    prob_n=getprob(par,1,1);
    prob_a=getprob(par,r_a,r_a);
    temp1_n=sum(prob_n[0:(out+1)]);
    temp2_n=sum(prob_n[0:out]);
    temp1_a=sum(prob_a[0:(out+1)]);
    temp2_a=sum(prob_a[0:out]);
    result=beta*(temp1_n+temp2_n-temp1_a-temp2_a);
    return(result);
}

f_W_t=function(w,out,r0,r1,r_a){
    p_t_1=0;
    p_t_2=100;
    w_t_1=getw(p_t_1,out,r_a);
    w_t_2=getw(p_t_2,out,r_a);
    w_l=min(w_t_1,w_t_2);
    w_u=max(w_t_1,w_t_2);
if (r_a < 1) {
  if (w <= w_l | w >= w_u) return(0)
  else { p_t_c = getw_I(w, out, r_a)
    result = -getpi(p_t_c, out, r0, r1)*f_p(p_t_c)/dw(p_t_c, out, r_a);
    return(result);
  }
}

else {
  if (w <= w_l | w >= w_u) return(0)
  else { p_t_c = getw_I(w, out, r_a)
    result = getpi(p_t_c, out, r0, r1)*f_p(p_t_c)/dw(p_t_c, out, r_a);
    return(result);
  }
}

f_W = function(w, r0, r1, r_a) {
  J = 3;
  temp = rep(0, J);
  for (i in 1:J) temp[i] = f_W_t(w, i - 1, r0, r1, r_a);
  f_W = sum(temp);
  return(f_W);
}

# Chebyshev polynomials
T = function(z, j, h) {
  t = cos((j - 1)*acos((2*z - h)/h));
  return(t)
}

# Nodes
z = function(i, h, N) {

```r
z=h/2*(1+cos((2*i-1)*pi/(2*N)));
return(z)
}

nodes=function(h,N){
  z=numeric(N);
  for(i in 1:N) z[i]=z(i,h,N)
  return(z)
}

#############Gaussian quadrature####################################
integ_t=function(out,h,j,loc,r0,r1,r_a,N_Gaussian){
  if(N_Gaussian==40){
    p=numeric(20);
    p[1]=0.9982377097105592; p[2]=0.9907262386994570;
    p[3]=0.9772599499837742; p[4]=0.9579168192137916;
    p[5]=0.9328128082786765; p[6]=0.9020988069688742;
    p[7]=0.8659595032122595; p[8]=0.8246122308333116;
    p[9]=0.7783056514265194; p[10]=0.7273182551899271;
    p[11]=0.6719566846141795; p[12]=0.6125538896679802;
    p[13]=0.5494671250951282; p[14]=0.4830758016861787;
    p[15]=0.4137792043716050; p[16]=0.3419940908257584;
    p[17]=0.2681521850072536; p[18]=0.1926975807013710;
    p[19]=0.1160840706752552; p[20]=0.0387724175060508;
    p=c(-p,rev(p));
    w=numeric(20);
    w[1]=0.004521277097105592; w[2]=0.01049828453115281;
    w[3]=0.0164210583190788; w[4]=0.02224584919416695;
    w[5]=0.02793700698002340; w[6]=0.03346019528254784;
    w[7]=0.03878216797447201; w[8]=0.04387090818567327;
    w[9]=0.04869580763507223; w[10]=0.05322784698393682;
    w[11]=0.05743976909939155; w[12]=0.06130624249292893;
    w[13]=0.06480401345660103; w[14]=0.0679120458152390;
    w=c(-w,rev(w));
  }
```

81
Risk-Adjusted CUSUM Charts and SMR

Tang Xu

w[15]=0.07061164739128677; w[16]=0.07288658239580405;
326 w[17]=0.07472316905796826; w[18]=0.07611036190062624;
327 w[19]=0.07703981816424796; w[20]=0.07750594797842481;
328 w=c(w,rev(w));
329 }
330 if(N_Gaussian==80){
331 p=numeric(40);
332 p[1]=0.019511383256793997654; p[2]=0.058504437152420668629;
333 p[3]=0.09740839841584599063; p[4]=0.136164022809143886559;
334 p[5]=0.174712291832646812559; p[6]=0.212994502857666132572;
335 p[7]=0.2509523589227120493; p[8]=0.288528054884511853109;
336 p[9]=0.325664370747701914619; p[10]=0.362304753499487315619;
337 p[11]=0.398393405881969227024; p[12]=0.433875370831756093062;
338 p[13]=0.468696615170544477036; p[14]=0.502804111888784987594;
339 p[15]=0.536145920897131932020; p[16]=0.568671268122709784725;
340 p[17]=0.600330622829751743155; p[18]=0.63107573046871966248;
341 p[19]=0.660859898986119801736; p[20]=0.689637644342027600771;
342 p[21]=0.717365185362099880254; p[22]=0.74400029758359727317;
343 p[23]=0.769502420135041373866; p[24]=0.793832717504605449949;
344 p[25]=0.816954138681463470371; p[26]=0.838831473580255275617;
345 p[27]=0.859431406663111096977; p[28]=0.878722567678213828704;
346 p[29]=0.896675579438770683194; p[30]=0.91326310257175654165;
347 p[31]=0.928459877172445795953; p[32]=0.942242761309872674752;
348 p[33]=0.954590766343634905493; p[34]=0.965485089043799251452;
349 p[35]=0.974909140585727793386; p[36]=0.982849572738629070418;
350 p[37]=0.989291302499755531027; p[38]=0.994227540965688277829;
351 p[39]=0.997649864398237688900; p[40]=0.999553822651630629880;
352 p=c(-rev(p),p);
353 w=numeric(40);
354 w[1]=0.039017813656306654811; w[2]=0.03895839562769531199;
355 w[3]=0.038839651059051968932; w[4]=0.038661759774076463327;
356 w[5]=0.038424993006959423185; w[6]=0.03812971134477638344;
357 w[7]=0.037776364362001397490; w[8]=0.037365490238730490027;


360 \[ w[9]=0.036897714638276008839; \]
361 \[ w[10]=0.035160529044747593496; \]
362 \[ w[11]=0.03473120451753928794; \]
363 \[ w[12]=0.033294193953416054603; \]
364 \[ w[13]=0.031710174188114701642; \]
365 \[ w[14]=0.03027232175959780661; \]
366 \[ w[15]=0.028259816057267682397; \]
367 \[ w[16]=0.0260752357655117903; \]
368 \[ w[17]=0.02492253576415491105; \]
369 \[ w[18]=0.023731882865930101293; \]
370 \[ w[19]=0.022505090246332461926; \]
371 \[ w[20]=0.02124402611578206399; \]
372 \[ w[21]=0.019950610878141998929; \]
373 \[ w[22]=0.018626814208299031429; \]
374 \[ w[23]=0.017274652056269306359; \]
375 \[ w[24]=0.015896183583725688042; \]
376 \[ w[25]=0.014493508040509076117; \]
377 \[ w[26]=0.013068761592401339294; \]
378 \[ w[27]=0.011624114120797826916; \]
379 \[ w[28]=0.010161766041103064521; \]
380 \[ w[29]=0.008683945269260858426; \]
381 \[ w[30]=0.007192904768117312753; \]
382 \[ w[31]=0.005690922451403198649; \]
383 \[ w[32]=0.004180313124694895237; \]
384 \[ w[33]=0.002663533589512861669; \]
385 \[ w[34]=0.001144950003186941534; \]
386 \[ w=c(rev(w),w); \]
387 
388 \( p_t_1=0; \)
389 \( p_t_2=100; \)
390 \( w_t_1=getw(p_t_1, out, r_a); \)
391 \( w_t_2=getw(p_t_2, out, r_a); \)
392 \( w_l=min(w_t_1, w_t_2); \)
393 \( w_u=max(w_t_1, w_t_2); \)
394 
395 \[ int_l=0; int_u=h; \]
396 \[ ll=w_l; ul=w_u; \]
397 \[ int_l_n=max(int_l, ll+loc); \]
398 \[ int_u_n=min(int_u, ul+loc); \]
399 
400 \if(int_l_n>=int_u_n) \{ \)
401 \[ p_n=(int_u_n-int_l_n)/2*(p-1)+int_u_n; \]
402 \[ fp=numeric(N_Gaussian); \]
403 \for(i in 1:N_Gaussian) \{ \)
404 \[ fp[i]=T(p_n[i], j, h)*f_W_t(p_n[i]-loc, out, r0, r1, r_a) \}
405 
406 { \)
407 }
396  int=(int_u_n-int_l_n)/2*(w%*%fp);
397  
398  int=as.numeric(int);
399  return(int);
400  }
401
402
403  integ_use=function(h,j,loc,r0,r1,r_a,N_Gaussian){
404    J=3;
405    temp=rep(0,3);
406    for(i in 1:J) temp[i]=integ_t(i-1,h,j,loc,r0,r1,r_a,N_Gaussian);
407    result=sum(temp);
408    return(result);
409  }
410
411  getmatrix=function(h,r0,r1,r_a,N,N_Gaussian_1,N_Gaussian_2){
412    z=nodes(h,N);
413    m=matrix(0,N,N);
414
415    for(i in 1:N)
416      for(j in 1:N){
417        m[i,j]=T(0,j,h)*F_W(-z[i],r0,r1,r_a,N_Gaussian_1)+
418          integ_use(h,j,z[i],r0,r1,r_a,N_Gaussian_2)-T(z[i],j,h);
419      }
420    return(m)
421  }
422
423  getc=function(h,r0,r1,r_a,N,N_Gaussian_1,N_Gaussian_2){
424    de=getmatrix(h,r0,r1,r_a,N,N_Gaussian_1,N_Gaussian_2);
425    b=rep(-1,N);
426    c=solve(de,b);
427    return(c)
428  }
429
430  getarl=function(startpoint,h,r0,r1,r_a,N,N_Gaussian_1,N_Gaussian_2){
431    u=startpoint;
c=getc(h,r0,r1,r_a,N,N_Gaussian_1,N_Gaussian_2);

p_t=rep(0,N);

for(j in 1:N) p_t[j]=T(u,j,h);

arl=c%*%p_t;

arl=as.numeric(arl);

return(arl)
References


Special Issue in Honour of Professor Shelemyahu Zacks, 9, 3-21.


Chapter 2. Standardized Mortality Ratio

Abstract

Measuring quality of medical practice is becoming increasingly prominent in quality management because it is a key component in improving quality and efficiency in health care. The traditional standardized mortality ratio (SMR) compares the mortality rate of a study population with that of a reference population by comparing the actual number of deaths in the study population with the expected number of deaths in the study population assuming the mortality rate of the reference population. In order to measure the performance of a surgeon or a group of surgeons in a hospital in performing a particular type of surgical operation, the SMR is used to compare the observed number of deaths in a sample with an estimated number of deaths usually calculated based on the average performance of a group of surgeons in a hospital, region, country or certain parts of the world. The estimated number of deaths in a sample associated with this new application of SMR is not a constant but a random variable. This means that all existing results for the traditional SMR may no longer be valid for the new SMR. In this chapter, the asymptotic distribution of the SMR based on an estimated number of deaths are derived. We also use the bootstrap procedure to estimate the finite-sample distribution. Both type I errors and powers of confidence intervals constructed using the asymptotic and bootstrap distributions of SMR are investigated and compared with existing methods.

KEY WORDS: Bootstrap; Cardiac operations; Euroscores; Hospital standardized mortality ratio; Logistic regression; Odds ratio; Parsonnet scores; Patient mix; Quality monitoring; risk distributions.
§1. Introduction

The standardized mortality ratio (SMR) is usually defined as the ratio $N_S/N_R$ where $N_S$ is the number of deaths in a study population and $N_R$ is the expected number of deaths in the study population assuming the mortality rate of a reference population. Hence, the SMR allows the mortality rate in a study population to be compared with that of a reference population. According to Keiding (1987), Dale (1777) was probably the earliest to explain and use the SMR. The SMR is described in many biostatistics textbooks, for example Pagano and Gauvreau (2000, pages 66–95), Forthofer et al. (2007, page 55), Armitage et al. (2008, pages 659–666), Antonisamy et al. (2010, pages 241–252) and Rosner (2011, pages 253–256). It is frequently used in various fields including occupational epidemiology (Ulm, 1989; Reid et al., 2008; Mirabelli et al., 2008), cancer studies (Breslow and Day, 1987; Mastrangelo et al., 2008; Gun et al., 2008; Guha, 2010; Bagary, 2011), heart studies (Humblet et al., 2008; Hickey et al., 2011; Nieuwkamp et al., 2011) and medical care (Thomas and Hofer, 1999; Sonnenberg, 2008; Zeger et al., 2008; Gensberg et al., 2009; Steenland et al., 2010; Beard et al., 2011).

Quite often in practice, instead of an entire study population, a sample is taken from the study population. The SMR is then defined as the ratio $O/E$ where the statistic $O$ now represents the observed number of deaths in the sample and the quantity $E$ represents the expected number of deaths in the sample assuming the mortality rate of a reference population. In order to derive a confidence interval for the SMR, the random variable $O$ is usually assumed to follow a Poisson distribution (Vandenbroucke, 1982; Ury, 1985 and Ulm, 1989). If the sample size is large and the mortality rate is small, the number of deaths can be approximated using a Poisson random variable. An approximate confidence interval for the SMR is obtained by first deriving a confidence interval for the expectation of $O$, and then dividing the interval with $E$. Bartlett (1936) showed that if $X$ is a Poisson random variable with mean $\lambda$, then $\sqrt{X}$ is an approximately normal random variable with mean
\[ \lambda \] and variance \( 1/4 \). Using this result, Vandenbroucke (1982) obtained a 95% confidence interval for the Poisson mean as \((\sqrt{\bar{O}} - 1)^2, (\sqrt{\bar{O}} + 1)^2\). In general, Vandenbroucke’s \((1 - \alpha)100\%\) confidence interval for the Poisson mean is given as \((\sqrt{\bar{O}} - z_{\alpha/2}/2)^2, (\sqrt{\bar{O}} + z_{\alpha/2}/2)^2\) where \(z_{\alpha/2}\) is the \(100(1 - \alpha/2)\)th standard normal percentile. Ury (1985) considered normal approximation of \(O\) instead of \(\lambda\) and with simple correction factors. His 95% and 99% confidence intervals for the Poisson mean are given as \((O - 1.96\sqrt{O} + 1, O + 1.96\sqrt{O} + 2)\) and \((O - 2.58\sqrt{O} + 2, O + 2.58\sqrt{O} + 3)\) respectively.

Clopper-Pearson (1934) provided a conservative approach of constructing a confidence interval and illustrated the approach for the binomial proportion. Suppose \(X\) is a Poisson random variable with mean \(\lambda\). If \(O = o\) is observed, the Clopper-Pearson interval is defined by \((\lambda_L, \lambda_U)\) where \(\lambda_L\) and \(\lambda_U\) are the solutions to the equations

\[
P(X \geq o | \lambda_L) = \frac{\alpha}{2} \quad \text{and} \quad P(X \leq o | \lambda_U) = \frac{\alpha}{2}.
\]

The Poisson probability can be calculated using a relationship between the Poisson and chi-square probabilities

\[
\sum_{i=0}^{m} e^{-\lambda} \lambda^i / i! = Pr(\chi^2_{2(m+1)} > 2\lambda).
\]

A proof of this relationship can be found in Johnson and Kotz (1969). It can be shown easily that the \(100(1 - \alpha)\%\) Clopper-Pearson confidence limits \(\lambda_L\) and \(\lambda_U\) based on \(O\) can be determined from the equations

\[
Pr(\chi^2_{2O} \leq 2\lambda_L) = \frac{\alpha}{2} \quad \text{and} \quad Pr(\chi^2_{2(O+1)} \leq 2\lambda_U) = 1 - \frac{\alpha}{2}.
\]

Ulm (1989) describe the use of these control limits in finding the confidence interval of SMR. The chi-square probability can also be approximated using the Wilson-Hilferty approximation as

\[
Pr(\chi^2_{2(m+1)} > 2\lambda) \approx Pr\left(Z > 3 \left[ \left( \frac{\lambda}{m+1} \right)^{1/3} - 1 + \frac{1}{9(m+1)} \right] \sqrt{m+1} \right),
\]
where $Z$ is the standard normal random variable. Pearson and Hartley (1958) tabulated the Clopper-Pearson confidence interval for the Poisson mean from 0 to 50. Using Pearson and Hartley’s table and the exact relationship between the Poisson and chi-square probabilities, Bailer and Ederer (1964) provided a table for determining the Clopper-Pearson confidence interval for the SMR for $O$ from 1 to 50. For $O$ greater than 50, they used the Wilson-Hilferty approximation to the chi-square probability to approximate the Clopper-Pearson confidence interval for the SMR. Byar (Breslow and Day, 1987) provided an explicit formula for such a confidence interval:

$$
\left( \left[ 1 - \frac{1}{9O} - \frac{z_{\alpha/2}}{3\sqrt{O}} \right]^3 \frac{O}{E}, \left[ 1 - \frac{1}{9(O+1)} + \frac{z_{\alpha/2}}{3\sqrt{O+1}} \right]^3 \frac{O+1}{E} \right).
$$

Descriptions for constructing the Clopper-Pearson confidence interval for the SMR can also be found in Mulder (1983) and Ulm (1989). Some authors called this an ‘exact’ confidence interval because it is calculated using the Poisson distribution but Clopper and Pearson (1934) made it clear that the actual coverage probability is equal to or greater than the nominal confidence level. The discreteness of the Poisson random variable means that exact confidence interval is usually not possible even if the Poisson distribution is used in deriving the interval. It should also be noted that the Poisson assumption may not be true.

If there are $k$ different mortality rates for $k$ groups of a study population defined by age for example (see Kielding, 1987), the SMR can be written as

$$
\text{SMR} = \frac{N_S}{N_R} = \frac{\sum_{i=1}^{k} N_i \alpha_i}{\sum_{i=1}^{k} N_i \lambda_i},
$$

where $N_i$ is subpopulation size of the $i$th group in the study population, $\alpha_i$ and $\lambda_i$ are the mortality rates of the $i$th group of the study and reference populations respectively. If a random sample is taken from the study population, then the
resulting SMR is given by
\[ \text{SMR} = \frac{O}{\sum_{i=1}^{k} n_i \lambda_i}, \]
where \( O \) is the observed number of deaths in the sample and \( n_i \) is the subsample size of the \( i \)th group in the sample. Note that the denominator is not a constant but a random variable determined by the random subsample sizes in a sample. It can be shown that
\[ E\left( \sum_{i=1}^{k} n_i \lambda_i \right) = \frac{n}{N} \sum_{i=1}^{k} N_i \lambda_i \]
where \( n = n_1 + \ldots + n_k \) and \( N = N_1 + \ldots + N_k \). This shows that \( \sum_{i=1}^{k} n_i \lambda_i \) is an estimate of the expected number of deaths in a sample of size \( n \). The statistical properties of this SMR remain unclear.

Measuring quality of medical practice is becoming increasingly prominent in quality management because it is a key component in improving quality and efficiency in health care. The hospital standardized mortality ratio (HSMR) was developed by Jarman et al. (1999) for measuring performances of hospitals. According to the HSMR Technical Notes published by the Canadian Institute for Health Information (2012), the HSMR is defined as the observed number of deaths in a hospital divided by the expected number of deaths based on 65 diagnostic groups which account for 80% of in-hospital deaths, excluding patients who received palliative care. If the risk distributions of patients remain the same, then the HSMR tracked over time indicates how successful hospitals and health regions have been in reducing in-hospital deaths, and this could lead to improved medical care. In USA, the Institute of Health Improvement (Whittington, 2005) is using the HSMR in their campaigns to improve the safety of patients by implementing strategies to reduce mortality. In England, the National Patient Safety Agency (Thomson, 2007) has adopted the HSMR as a high level track measure for patient safety. The Canadian Institute for Health Information has led the effort in promoting the use of HSMR for Canada and publishes results for eligible facilities and regions in all provinces outside Quebec.

Although the term ‘expected’ is used in the definition of HSMR, it is clear from
the way that the ‘expected’ number of deaths is defined and calculated according to Appendix I of the HSMR Technical Notes published by the Canadian Institute for Health Information (2012), it is an estimated number of deaths. Thus, existing confidence intervals derived for the SMR including the Byar’s interval (see for examples, Guidelines for Using and Developing Rates for Public Health Assessment, 2002 and HSMR Technical Notes published by the Canadian Institute for Health Information, 2012), are no longer valid for the HSMR unless \( \sum_{i=1}^{k} n_i \lambda_i \) is a good approximation of its expected quantity \( \frac{n}{N} \sum_{i=1}^{k} N_i \lambda_i \). Faris et al. (2003) showed using simulation that assuming \( \sum_{i=1}^{k} n_i \lambda_i \) is a constant will result in a wider confidence interval for the SMR.

In this chapter, we consider \( k \) different mortality rates for \( k \) groups of a study population and study the statistical properties of the SMR based on a random sample taken from the study population. In Section 2, the statistical properties and asymptotic distribution of the SMR are investigated. In Section 3, we develop confidence intervals for the SMR based on the asymptotic and bootstrap distributions of SMR. The coverage probabilities of these confidence intervals and those based on existing methods (in which the denominator of the SMR, \( \sum_{i=1}^{k} n_i \lambda_i \) is assumed to be a constant) are also studied. The powers of the various confidence intervals are then investigated in Section 4. The SMR is highly dependent on the patients’ risk distribution and this relationship is investigated in Section 5. Real data sets are used to illustrate the various confidence intervals in Section 6. A conclusion is given in Section 7.

\[ \text{2. Statistical Properties of SMR} \]

\[ \text{2.1 Definition and Expectation of SMR} \]

Suppose a patient is to undergo a cardiac surgery. The surgical outcome of the patient can be represented by \( Y \) which is 1 if the patient dies within 30 days and 0 otherwise. The patient’s conditions like age, blood pressure, existence of certain
disease like diabetes, morbid obesity et al. will be determined and this information can be summarized as a Parsonnet score (Parsonnet et al., 1989). The Parsonnet score is an integer from 0 to 100 and it is a measure of the risk of death of a patient who undergoes a cardiac surgery. A patient’s Parsonnet score $S$ can then be used to estimate the probability of death $X$ by using a logistic regression model $m(\cdot)$ built using some past surgical data as $X = m(S)$. The quantity $X$ is an estimate of the probability of death resulting from an operation assuming the average performance of all surgeons in the data. The model defines a one-to-one relationship between $S$ and $X$. For a patient who is operated on by a particular surgeon, let $d(X)$ be the true risk of death and define $D(S) = d(m(S))$. This means that given a patient with Parsonnet score $S$, the true risk of death operated on by the surgeon will be represented by $D(S)$. If a model is fitted using a different data set, $m(\cdot)$ and $d(\cdot)$ will change but $D(\cdot) = d(m(\cdot))$ will remain the same. In effect, we have defined $d(\cdot)$ and $D(\cdot)$ as measures of performance of a surgeon. The main difference between $d(\cdot)$ and $D(\cdot)$ is that $d(\cdot)$ is model dependent whereas $D(\cdot)$ is not. A surgeon who performs uniformly better than a surgeon with an average performance will have $D(s) < m(s)$ for all $s$ or $d(x) < x$ for all $x$. Similarly, a surgeon who performs uniformly worse than average will have $D(s) > m(s)$ for all $s$ or $d(x) > x$ for all $x$.

Let $S_1, S_2, \cdots, S_n$ be the Parsonnet scores of a random sample of $n$ patients. The SMR is defined as

$$\text{SMR} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i},$$

where $\sum_{i=1}^{n} Y_i$ is the observed number of deaths and $\sum_{i=1}^{n} X_i$ is an estimate of the expected number of deaths in a sample of size $n$ given the model $x = m(s)$. The expected number of deaths in the sample of size $n$ is calculated based on the average performance of all the surgeons whose data were used to fit the model $x = m(s)$. A model $x = m(s)$ that underestimates the risk $x$ will inflate the SMR. Similarly, a model that overestimates the risk $x$ will deflate the SMR. The SMR is thus dependent on how the risk $x$ is estimated and hence it is subjected to possible
abuses or manipulations.

Given the estimated risk $X = x$ of a patient, $Y$ is a Bernoulli random variable with probability of death from an operation $d(x)$. Note that the SMR is a ratio of two random variables and using Taylor’s approximation, it can be shown that

$$E(\text{SMR}) \approx \frac{\mu_Y}{\mu_X},$$

(2)

where $\mu_X = \sum xp(x)$, $\mu_Y = \sum d(x)p(x)$ and $p(\cdot)$ is the probability mass function (pmf) of $X$. Thus, the SMR is an almost unbiased estimator of $\frac{\mu_Y}{\mu_X}$. The quantity $\frac{\mu_Y}{\mu_X}$ can also be expressed as

$$\frac{\mu_Y}{\mu_X} = \sum \frac{d(x)}{x} \cdot \frac{xp(x)}{\mu_X},$$

(3)

which can be viewed as a weighted average of the performance measure $d(x)/x$ with weight $xp(x)/\mu_X$. The weight depends on both $x$ and $p(x)$, hence $\frac{\mu_Y}{\mu_X}$ is influenced by both the estimated risk of death and its distribution. Equation (3) also indicates that for two equally competent surgeons with the same $d(\cdot)$, their SMRs can be different, depending on the risk distribution of their patients. Therefore, a direct comparison of the magnitudes of SMRs may not be meaningful when their patients’ risk distributions are different.

2.2 Asymptotic Distribution of SMR

We will derive the asymptotic distribution of SMR.

Theorem 1. Consider a random sample of $n$ patients with their probabilities of death and surgical outcomes denoted as $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, then the SMR $= \frac{\sum X_i}{\sum Y_i}$ has the following asymptotic distribution

$$\sqrt{n}(\text{SMR} - \frac{\mu_Y}{\mu_X}) \overset{L}{\rightarrow} N(0, \frac{\mu_Y^2}{\mu_X^2} \sigma_X^2 - 2\frac{\mu_Y}{\mu_X^2} \sigma_{XY} + \frac{1}{\mu_X^2} \sigma_Y^2),$$

where $\mu_X = E(X), \mu_Y = E(Y), \sigma_X^2 = \text{Var}(X), \sigma_Y^2 = \text{Var}(Y)$ and $\sigma_{XY} = \text{Cov}(X,Y)$. 

96
Proof. Let $W_i = (X_i, Y_i)^T$ and note that $W_1, W_2, \ldots, W_n$ are independent and identically distributed (iid) random vectors. The expectation and covariance matrix of $W_i$ are given as $E(W_i) = (\mu_X, \mu_Y)^T$ and

$$
\text{Var}(W_i) = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}.
$$

Let $f(x, y) = y/x$, then $\text{SMR} = f(\bar{X}, \bar{Y})$. Also, define $\nabla^T f(\mu_X, \mu_Y) = (-\mu_Y/\mu_X^2, 1/\mu_X)$. Using Delta method, it can be shown that

$$
\sqrt{n}(f(X, Y) - f(\mu_X, \mu_Y)) \xrightarrow{\text{d}} N(0, \tau^2),
$$

where $\tau^2 = \nabla^T f(\mu_X, \mu_Y) \text{Var}(W_i) \nabla f(\mu_X, \mu_Y) = \frac{\mu_Y^2}{\mu_X^2} \sigma_X^2 - \frac{2 \mu_Y}{\mu_X^2} \sigma_{XY} + \frac{1}{\mu_X^2} \sigma_Y^2$. This completes the proof.

§3. Comparison of Confidence Intervals of SMR

3.1 Some Basic Assumptions

Suppose the odds ratio of death associated with a surgeon is $Q = [d(x)/(1 - d(x))]/[x/(1 - x)]$ (Steiner et al. 2000). This means the odds of death of a patient operated on by the surgeon is $Q$ times the odds of death if the patient were to be operated on by a surgeon of average performance. A plot of $d(x)$ against $x$ for $Q = 0.5, 1$ and $2$ is displayed in Figure 3.1. For a surgeon with $Q = 1$, the true risk of death is represented by the diagonal line $d(x) = x$. Note that the true risk of death of a patient operated on by a surgeon with $Q > 1$ is larger than that of $Q = 1$. The opposite is true for $Q < 1$.

Consider a real data set of 6449 Parsonnet scores of cardiac patients and its histogram is shown in Figure 3.2. Each patient was operated on by a surgeon from a group of surgeons. This data set is denoted as population $P_0$. A real data set was used for the study so that any results obtained will be representative of real scenarios. A logistic regression model

$$
\log(x_0/(1 - x_0)) = \text{logit}(x_0) = \alpha_0 + \beta_0 s,
$$

(4)
is first fitted using $P_0$ and its associated surgical outcomes to obtain a relationship between the Parsonnet score $s$ and the risk of death $x$ assuming the average performance of all the surgeons. The two parameters were found to be $\alpha_0 = -3.605$ and $\beta_0 = 0.072$.

As an example of the covariance matrix $\text{Var}(W)$, we simulated 1000 samples which assume the risk profiles is the same as $P_0$ and the performance is characterized by $Q = 2$. The covariance matrix of this data will be given as

$$\text{Var}(W) = \begin{pmatrix} 0.0067 & 0.0094 \\ 0.0094 & 0.1140 \end{pmatrix}.$$
Figure 3.1. Plots of true probability of death $d(x)$ of a patient operated on by a surgeon characterized by the odds ratio of death $Q$ against the probability of death $x$ of a patient operated on by a surgeon with an average performance.

Figure 3.2. The frequency distribution of Parsonnet scores of a group of 6449 patients who underwent cardiac operations.
### 3.2 Methods for Constructing Confidence Intervals

Suppose that the observed data for a surgeon is given by \((S_1, Y_1), (S_2, Y_2), \ldots, (S_n, Y_n)\), where \(S_i\) is the Parsonnet score of the \(i\)th patient and \(Y_i\) is the surgical outcome. The observed number of deaths is given as \(O = \sum_{i=1}^{n} Y_i\). The probability of death from an operation can be estimated using a logistic model \(X_i = m(S_i)\). The estimated number of deaths in the sample is then given as \(\sum_{i=1}^{n} X_i\). There are basically 4 approaches of constructing confidence intervals for the SMR of which the first two are existing procedures.

1. Normal approximation of the Poisson distribution: Vandenbroucke’s and Ury’s confidence intervals;
2. Clopper-Pearson approach based on the Poisson distribution: Ulm’s and Byar’s confidence intervals;
3. Confidence intervals based on the asymptotic normal theory;

The first two types assume that (i) the numerator \(O\) is distributed as a Poisson random variable, and (ii) the denominator of the SMR, \(\sum_{i=1}^{n} X_i\) is a constant. The first two types differ in the types of normal approximation used for the Poisson distribution. The other two types do not make such assumptions and they use either the asymptotic normal distribution or bootstrap approach in estimating the percentiles of the SMR.

### 3.3 Comparison of Coverage Probabilities

In order to investigate the coverage probabilities of the confidence intervals, we perform a simulation study. The logistic regression model in equation (4) is first fitted using \(P_0\) and its associated surgical outcomes to obtain a relationship between the Parsonnet score \(s\) and the risk of death \(x_0\) assuming the average performance of all the surgeons. In our simulation study, we will assume that the risk of death calculated using this model is the true risk of death of a patient given a Parsonnet score. We consider taking random samples of Parsonnet scores from
population $P_1$ (which is the same as population $P_0$) and the true risks of death can be determined using the model in equation (4) for the simulation study. These risks will form the denominator of the SMR. Consider a surgeon $A$ with odds ratio of death $Q_A$, operating on patients. In order to construct a confidence interval for the SMR of surgeon $A$, we will calculate the true risks of death for patients using $Q_A = [d(x)/(1 - d(x))]/[x_0/(1 - x_0)]$ so as to simulate the surgical outcomes which will form the numerator of the SMR. Each SMR is simulated 1000 times for $Q_A = 0.5, 0.6, ..., 2.0$. The various confidence intervals are then calculated to determine their respective coverage probabilities. This is done by checking whether a confidence interval covers $\mu_Y/\mu_X$ or not. The calculation of $\mu_Y/\mu_X$ is described in equation (3).

Figure 3.3 shows the coverage probabilities of the 4 types of confidence intervals. There are clear differences in their coverage probabilities. As expected, Ulm’s and Byar’s coverage probabilities are almost identical because both are based on the same Clopper-Pearson approach and differ only in the calculation of the Poisson probabilities. The coverage probability is always above the 0.95 level as dictated by the Clopper-Pearson approach. Their coverage probabilities are around 0.975 and they are the most conservative. Ury’s confidence interval is nearly as conservative as Ulm’s and Byar’s confidence intervals. Vandenbroucke’s confidence interval is slightly less conservative with its coverage probability hovers around 0.969. It is clear from Figure 3.3 that the first 2 types of confidence intervals are conservative. The confidence interval based on the asymptotic normal theory is the best in terms of coverage probability which hovers around 0.95. This shows that the normal approximation is adequate. The confidence interval based on the bootstrap approach is similar to that based on asymptotic normal theory with slightly smaller coverage probabilities.

In order to study the influence of the risk distribution of patients operated on by a surgeon on the coverage probabilities, we repeat the simulation study by con-
considering 3 other risk distributions that are different from \( P_1 \). The risk distributions of the 3 populations are summarized in Table 3.1. The risk distribution \( P_2 \) has more low-risk patients than \( P_1 \). The other two have more high-risk patients. Risk distribution \( P_4 \) is quite different from \( P_1 \) and it is representative of patients seeking treatments from a surgeon or hospital that treats mostly high-risk patients. Figures 3.4–3.6 show the coverage probabilities for the 3 risk distributions. Table 3.2 shows that average coverage probabilities (over the range of \( Q \)) of the various types of confidence intervals with respect to the risk distributions \( P_1–P_4 \). Figures 3.4–3.6 and Table 3.2 show that the confidence intervals based on Poisson approximation remain conservative, while those based on asymptotic normal theory and bootstrap approach have coverage probabilities average to around 0.95. As the proportion of high-risk patients increases from \( P_1 \) to \( P_4 \), the confidence intervals based on Poisson approximation become even more conservative for \( Q \) greater than 1 because the Poisson approximation deteriorates as the mean number of deaths increases.
Table 3.1. Risk distributions of the 5 populations of Parsonnet scores

<table>
<thead>
<tr>
<th>Population</th>
<th>Low</th>
<th>Elevated</th>
<th>Significantly elevated</th>
<th>High</th>
<th>Very high</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_0$</td>
<td>37%</td>
<td>24%</td>
<td>15%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>$P_1$</td>
<td>37%</td>
<td>24%</td>
<td>15%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>$P_2$</td>
<td>47%</td>
<td>34%</td>
<td>15%</td>
<td>1%</td>
<td>3%</td>
</tr>
<tr>
<td>$P_3$</td>
<td>32%</td>
<td>19%</td>
<td>15%</td>
<td>16%</td>
<td>18%</td>
</tr>
<tr>
<td>$P_4$</td>
<td>3%</td>
<td>1%</td>
<td>15%</td>
<td>34%</td>
<td>47%</td>
</tr>
</tbody>
</table>

Table 3.2. Average of coverage probabilities of various confidence intervals under different risk distributions

<table>
<thead>
<tr>
<th>Population</th>
<th>Vandenbroucke</th>
<th>Ury</th>
<th>Ulm</th>
<th>Byar</th>
<th>Asymptotic</th>
<th>Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>0.969</td>
<td>0.974</td>
<td>0.975</td>
<td>0.975</td>
<td>0.949</td>
<td>0.946</td>
</tr>
<tr>
<td>$P_2$</td>
<td>0.959</td>
<td>0.968</td>
<td>0.968</td>
<td>0.968</td>
<td>0.946</td>
<td>0.941</td>
</tr>
<tr>
<td>$P_3$</td>
<td>0.972</td>
<td>0.976</td>
<td>0.976</td>
<td>0.976</td>
<td>0.947</td>
<td>0.945</td>
</tr>
<tr>
<td>$P_4$</td>
<td>0.975</td>
<td>0.978</td>
<td>0.978</td>
<td>0.978</td>
<td>0.948</td>
<td>0.945</td>
</tr>
</tbody>
</table>
Figure 3.3. The simulated coverage probabilities of various confidence intervals when the underlying risk distribution is $P_1$.

Figure 3.4. The simulated coverage probabilities of various confidence intervals when the underlying risk distribution is $P_2$. 

104
Figure 3.5. The simulated coverage probabilities of various confidence intervals when the underlying risk distribution is $P_3$.

Figure 3.6. The simulated coverage probabilities of various confidence intervals when the underlying risk distribution is $P_4$. 
§4. Power of the 4 Test Procedures

Consider a surgeon with odds ratio of death $Q$ operating on patients. We want to test the hypothesis $H_0 : \mu_Y/\mu_X = \sum d_0(x)p(x)/\sum xp(x)$ versus $H_1 : \mu_Y/\mu_X \neq \sum d_0(x)p(x)/\sum xp(x)$, where $x$ is the true risk of death of the $i$th patient assuming the average performance of surgeons which is calculated using model (4) and $d_0(x)$ is defined by $d_0(x)/(1 - d_0(x)) = Q_0x/(1 - x)$. We consider $Q_0 = 0.5, 1.0 \text{ and } 2.0$ in our simulation study. In order to investigate the power of the test procedures, we consider taking random samples of 1000 patients from population $P_1$. The true risk of death of the $i$th patient assuming the average performance of surgeons is calculated as $x_i$ using model (4). We investigate the power of the 4 types of procedures for $Q/Q_0 = 0.5, 0.6, \ldots, 2.0$. The true risk of death of the $i$th patient to be operated on by a surgeon with an associated $Q$ is calculated as $d(x_i)$. The outcome of the surgical operation for the $i$th patient $y_i$ can be done by simulating a standard uniform random variate and compare it with $d(x_i)$. The various confidence intervals for the SMR are then calculated as described in Section 3. Based on 1000 samples simulated, the powers of the various confidence intervals can be obtained. The power curves of the 4 types of confidence intervals are shown in Figures 4.1–4.6 for (i) $Q_0 = 1, \alpha = 0.05$, (ii) $Q_0 = 1, \alpha = 0.01$, (iii) $Q_0 = 0.5, \alpha = 0.05$, (iv) $Q_0 = 0.5, \alpha = 0.01$, (v) $Q_0 = 2, \alpha = 0.05$, (vi) $Q_0 = 2, \alpha = 0.01$.

Figure 4.1 shows the power curves of the 6 procedures for $Q_0 = 1, \alpha = 0.05$. It can be seen that the power of the 6 procedures are approximately the same when $Q > Q_0$. However, the power of the asymptotic and bootstrap methods are significantly higher than the other methods when $Q < Q_0$. The difference in power is due to a difference between the normal and Poisson approximations. In order to assess the normality of SMR based on a sample size of 1000, we simulated 200 samples and calculated their SMRs. The resulting SMRs are then tested for normality using the 5% level of significance Shapiro-Wilk test. This is repeated 500 times and the proportion of times the test rejected the null hypothesis of normality
is found to be 0.084. This provides evidence that the normal approximation works well for the SMR based on a finite sample size of 1000. Figures 4.7–4.9 show the histograms of 1000 simulated SMRs based on $Q = 0.5$, 1 and 2 and samples taken from $P_1$. The exact asymptotic normal pdf of the SMR is also shown on each figure. Under the Poisson assumption, the distribution of the SMR can be approximated by the distribution of a Poisson random variable divided by $\mu_X$. This distribution is also shown on each figure. These figures show that the asymptotic normal distribution provides a good fit to the data. The Poisson approximation is right-skewed with a larger variance. This explains why the upper confidence limits of the SMR based on the Poisson approximation is larger and resulted in lower power because rejection is based on the upper limit when $Q < Q_0$. This is also confirmed with the results in Table 4.1 which shows the average of the differences (i) SMR – lower limit and (ii) upper limit – SMR, of the various types of 95% confidence intervals obtained by simulating 1000 SMRs based on $P_1$ and $Q = 1$. Table 4.1 shows that the average of the difference SMR – lower limit is very similar for all the confidence intervals. As for the average of the difference upper limit – SMR, the upper limits of SMR based on the Poisson approximation are larger.
Figure 4.1. The power of the test procedures using various confidence intervals for $Q_0 = 1, \alpha = 0.05$.

Figure 4.2. The power of the test procedures using various confidence intervals for $Q_0 = 1, \alpha = 0.01$. 
Figure 4.3. The power of the test procedures using various confidence intervals for $Q_0 = 0.5, \alpha = 0.05$.

Figure 4.4. The power of the test procedures using various confidence intervals for $Q_0 = 0.5, \alpha = 0.01$. 

109
Figure 4.5. The power of the test procedures using various confidence intervals for $Q_0 = 2, \alpha = 0.05$.

Figure 4.6. The power of the test procedures using various confidence intervals for $Q_0 = 2, \alpha = 0.05$. 

110
Figure 4.7. A histogram of 1000 simulated SMRs based on $Q = 0.5$ and random samples taken from $P_1$. The exact asymptotic normal pdf and the Poisson approximation of the SMR are displayed as solid and dotted lines respectively.

Figure 4.8. A histogram of 1000 simulated SMRs based on $Q = 1$ and random samples taken from $P_1$. The exact asymptotic normal pdf and the Poisson approximation of the SMR are displayed as solid and dotted lines respectively.
**Figure 4.9.** A histogram of 1000 simulated SMRs based on $Q = 2$ and random samples taken from $P_1$. The exact asymptotic normal pdf and the Poisson approximation of the SMR are displayed as solid and dotted lines respectively.

**Table 4.1.** The average lengths of (i) SMR – lower confidence limit and (ii) upper confidence limit – SMR of the various 95% confidence intervals obtained by simulating 1000 SMRs based on $P_1$ and $Q = 1$. The standard deviations are shown in the brackets.

<table>
<thead>
<tr>
<th></th>
<th>Van</th>
<th>Ury</th>
<th>Ulm</th>
<th>Byar</th>
<th>As</th>
<th>Bs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMR - Lower limit</td>
<td>0.226</td>
<td>0.225</td>
<td>0.225</td>
<td>0.225</td>
<td>0.220</td>
<td>0.225</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.013)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Upper limit -SMR</td>
<td>0.254</td>
<td>0.270</td>
<td>0.271</td>
<td>0.271</td>
<td>0.220</td>
<td>0.214</td>
</tr>
<tr>
<td></td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.014)</td>
<td>(0.013)</td>
<td>(0.015)</td>
</tr>
</tbody>
</table>
5. Effect of Risk Distributions on SMR

In this section, we will examine the effect of risk distribution of patients on the SMR. For the study, we consider random samples of Parsonnet scores taken with replacement from $P_0$. Samples of 100 Parsonnet scores are taken randomly from each of the five risk categories of $P_0$ as classified in National Adult Cardiac Surgical Database Report 2000-2001: low (0–4), elevated (5–9), significantly elevated (10–14), high (15–19), very high (20–100). The risk of death of a patient assuming the performance of an average surgeon, $x_0$ (and represented as $x$) is calculated using equation (4) and $x$ is assumed to be the true risk of death in the study. For each of the 5 samples, the patients are operated on by a surgeon with odds ratio of death $Q = 0.5$. In other words, the true risk of death of a patient operated on by the surgeon is given as $d(x) = Qx/(Qx+1-x)$. Since the pmf of $X$ is known for $P_0$, the quantity $\mu_Y/\mu_X$ ($E(SMR)$ is a close approximation of $\mu_Y/\mu_X$) can be calculated exactly using equation (2). This is repeated for a surgeon with $Q = 0.6, 0.7, \ldots, 2.0$.

Table 5.1 contains the quantity $\mu_Y/\mu_X$ classified by the risk category of patients and the odds ratio of death associated with a surgeon. The table reveals that the SMR of a surgeon with an average performance ($Q = 1$) is unaffected by the risk distribution. This is due to the fact that if $Q = 1$, then $d(x) = x$ and equation (2) yields $\mu_Y/\mu_X = 1$. For a surgeon who performs better than average ($Q < 1$), the SMR is larger if the risk is from a higher category. In contrast, for a surgeon who performs worse than average ($Q > 1$), the SMR is smaller if the risk is from a higher category. This is contrary to a popular misconception among hospital administrators and medical professionals that even risk-adjusted performance measures will fare worse if the patients operated on are of higher risks. For example, Kahn et al. (2007) commented that hospitals might shift the obligations of treating higher risk patients by transferring them out to other health care institutions in order to get a smaller SMR. The table shows that this is not true unless $Q$ is less than 1. For $Q$ closer to 1, the quantity $\mu_Y/\mu_X$ varies less across the risk categories. This shows
that the SMR does in fact adjust for the risks of patients. However, for a surgeon with $Q = 0.5$ for example, the SMR for the very high risk category can be 20% larger than that for a low risk category. Also, the SMR of a surgeon with $Q = 2.0$ operating on very high risk patients is smaller than the SMR of a surgeon with $Q = 1.8$ operating on low risk patients. Thus, it may not be meaningful to compare the magnitudes of two SMRs directly if their risk distributions are different.

<table>
<thead>
<tr>
<th>Risk category</th>
<th>Odds ratio of death associated with a surgeon, $Q$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td>Low</td>
<td>0.51</td>
</tr>
<tr>
<td>Elevated</td>
<td>0.51</td>
</tr>
<tr>
<td>Significantly elevated</td>
<td>0.52</td>
</tr>
<tr>
<td>High</td>
<td>0.52</td>
</tr>
<tr>
<td>Very high</td>
<td>0.60</td>
</tr>
</tbody>
</table>

In addition, we want examine the effects of the risk distributions on the confidence intervals. From the findings in Section 3, the confidence interval based on asymptotic normal distribution is the best among several methods in terms of coverage probability and power. Thus, we will use this confidence interval in the study. For a surgeon with $Q = 1$, consider the surgeon operating on patients taken from 4 populations $P_1$, $P_2$, $P_3$ and $P_4$. We set $P_1$ as the underlying standard population, and examine the effect of risk distributions with $P_2$ having more lower risk patients, and $P_3$ and $P_4$ having more higher risk patients. For each of the 4 populations, the quantity $\mu_Y/\mu_X$ ($E(SMR)$ is a close approximation of $\mu_Y/\mu_X$) and the standard deviation of SMR are calculated. In addition, proportions of 95% confidence intervals below and above $E(SMR)$ associated with $P_1$, are estimated based on 5000
random samples taken from each of the 4 populations. The results are displayed in Table 5.2. Since $Q = 1$, the quantity $E(SMR)$ is not affected by the change in the risk distribution. The proportion of 95% confidence interval of SMR below the underlying $E(SMR)$ does not change substantially when the risk distribution changes. This is also true for the proportion of 95% confidence intervals of SMR above the $E(SMR)$. These results imply that the effects of the risk distributions on the confidence interval of SMR is relatively small when $Q = 1$.

The study is repeated for $Q = 0.5$ and $Q = 2$. The results for $Q = 0.5$ are shown in Table 5.3. From the table, the proportion of 95% confidence intervals of SMR below the $E(SMR)$ associated with $P_1$ is much lower for $P_4$ than that for $P_1$. On the other hand, the proportion of 95% confidence intervals of SMR above the $E(SMR)$ associated with $P_1$ is much higher for $P_4$ than that for $P_1$. In other words, if the surgeon operates on more high risk patients, the surgeon will look worse in performance. From the same table, the proportion of 95% confidence intervals of SMR below the $E(SMR)$ associated with $P_1$ is much higher for $P_2$ than that for $P_1$. On the other hand, the proportion of 95% confidence intervals of SMR above the $E(SMR)$ associated with $P_1$ is much lower for $P_2$ than that for $P_1$. In other words, if the surgeon operates on more low risk patients, the surgeon will look better in performance. The results for $Q = 2$ are shown in Table 5.4. The tables shows that if the surgeon operates on more high risk patients, the surgeon will look better in performance. On the other hand, if the surgeon operates on more low risk patients, the surgeon will look worse in performance.

In summary, the results show that a worse than average surgeon can indeed “improve” his or her performance by operating on more high risk patients. Similarly, a better than average surgeon can “improve” his performance by operating on more low risk patients. This is contrary to popular believe that a surgeon can only “improve” his performance by operating on low risk patients (Kahn et al., 2007).
**Table 5.2.** (i) Expectation, (ii) standard deviation of SMR and (iii) estimated proportion of 95% confidence interval of SMR below or above $E$(SMR) associated with $P_1$, based on 5000 random samples taken from $P_2$, $P_1$, $P_3$ and $P_4$ for the case $Q = 1$

<table>
<thead>
<tr>
<th></th>
<th>$P_2$</th>
<th>$P_1$</th>
<th>$P_3$</th>
<th>$P_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E(SMR)</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>SD(SMR)</td>
<td>0.114</td>
<td>0.112</td>
<td>0.101</td>
<td>0.074</td>
</tr>
<tr>
<td>% of c.i. below 1.000</td>
<td>0.043</td>
<td>0.036</td>
<td>0.033</td>
<td>0.032</td>
</tr>
<tr>
<td>% of c.i. above 1.000</td>
<td>0.013</td>
<td>0.017</td>
<td>0.017</td>
<td>0.022</td>
</tr>
</tbody>
</table>

**Table 5.3.** (i) Expectation, (ii) standard deviation of SMR and (iii) estimated proportion of 95% confidence interval of SMR below or above $E$(SMR) associated with $P_1$, based on 5000 random samples taken from $P_2$, $P_1$, $P_3$ and $P_4$ for the case $Q = 0.5$

<table>
<thead>
<tr>
<th></th>
<th>$P_2$</th>
<th>$P_1$</th>
<th>$P_3$</th>
<th>$P_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E(SMR)</td>
<td>0.524</td>
<td>0.550</td>
<td>0.559</td>
<td>0.578</td>
</tr>
<tr>
<td>SD(SMR)</td>
<td>0.105</td>
<td>0.085</td>
<td>0.078</td>
<td>0.059</td>
</tr>
<tr>
<td>% of c.i. below 0.550</td>
<td>0.072</td>
<td>0.039</td>
<td>0.030</td>
<td>0.011</td>
</tr>
<tr>
<td>% of c.i. above 0.550</td>
<td>0.006</td>
<td>0.017</td>
<td>0.022</td>
<td>0.053</td>
</tr>
</tbody>
</table>

**Table 5.4.** (i) Expectation, (ii) standard deviation of SMR and (iii) estimated proportion of 95% confidence interval of SMR below or above $E$(SMR) associated with $P_1$, based on 5000 random samples taken from $P_2$, $P_1$, $P_3$ and $P_4$ for the case $Q = 2$

<table>
<thead>
<tr>
<th></th>
<th>$P_2$</th>
<th>$P_1$</th>
<th>$P_3$</th>
<th>$P_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>E(SMR)</td>
<td>1.866</td>
<td>1.758</td>
<td>1.722</td>
<td>1.643</td>
</tr>
<tr>
<td>SD(SMR)</td>
<td>0.192</td>
<td>0.129</td>
<td>0.129</td>
<td>0.090</td>
</tr>
<tr>
<td>% of c.i. below 1.758</td>
<td>0.009</td>
<td>0.035</td>
<td>0.060</td>
<td>0.232</td>
</tr>
<tr>
<td>% of c.i. above 1.758</td>
<td>0.061</td>
<td>0.016</td>
<td>0.010</td>
<td>0.001</td>
</tr>
</tbody>
</table>


6. Real Data Analysis

In this section, we will illustrate the methodology using a real data set. This real data comprises 6499 patients, operated on by 7 surgeons. In phase I, the data for the first 1906 patients are used to fit a logistic model,

\[
\text{logit}(x) = -3.6374 + 0.0748s.
\]

In phase II, the model is used to estimate the risk of death of a patient from an operation assuming the average performance of the 7 surgeons. The data for the other 4593 patients are then used to calculate the SMRs and confidence intervals for the 7 surgeons. There are less low-risk and more higher-risk patients in phase II than phase I.

Table 6.1 shows the SMRs and their 95% confidence intervals for the 7 surgeons. The number of patients operated on are also given. The lower limits are similar for all the intervals but the upper limits of the confidence intervals based on asymptotic normal and bootstrap distributions are closer to the SMR. In other words, the confidence interval based on asymptotic distribution and the bootstrap method are more accurate than the other methods. This difference is due to the right skewness of the Poisson approximation which was explained in the previous section. The results obtained here are consistent with our findings in the previous section.

Table 6.1 also shows that there is evidence that surgeons 1 and 2 have performances that are worse than the average and surgeon 6 has performance better than the average. There is no evidence to suggest that the other surgeons have performances significantly different from the average. The patients operated on by surgeon 2 are of higher risks than those of surgeon 1. This means surgeon 2’s performance could be worse if he were to operate on more low risk patients. Surgeon 6 is the only one who performed better than average. Most of his patients are of low or elevated risks and this might explain partly his good performance. The SMRs of surgeons 3 and 5 are found to be 0.805 and 0.744 respectively and they are only marginally non-significant. Note that risk distribution of patients operated on by
surgeon 5 is of a lower-risk distribution than that of surgeon 3. This means surgeon 3’s SMR could be lower if he were to operate on patients of a lower-risk distribution. Surgeon 7 is probably an average surgeon with a SMR of 0.99. Although surgeon 4’s SMR (1.351) is greater than surgeon 1’s SMR (1.223), there is not enough evidence to show that surgeon 4 is worse than average due to the small sample size. Note that the length of the confidence interval is about 0.4 for surgeon 1 and about 1.1 for surgeon 4. This is due to the small sample size associated with surgeon 4.
Table 6.1. Confidence intervals of 7 surgeons based on a real data set of 4593 patients

<table>
<thead>
<tr>
<th>Surgeon</th>
<th>Sample Size</th>
<th>Confidence Interval</th>
<th>Van</th>
<th>Ury</th>
<th>Ulm</th>
<th>Byar</th>
<th>As</th>
<th>Bs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1161</td>
<td>Low limit</td>
<td>1.004</td>
<td>1.004</td>
<td>1.004</td>
<td>1.004</td>
<td>1.009</td>
<td>1.004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>1.223</td>
<td>1.223</td>
<td>1.223</td>
<td>1.223</td>
<td>1.223</td>
<td>1.223</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>1.465</td>
<td>1.477</td>
<td>1.477</td>
<td>1.477</td>
<td>1.437</td>
<td>1.441</td>
</tr>
<tr>
<td>2</td>
<td>330</td>
<td>Low limit</td>
<td>1.032</td>
<td>1.032</td>
<td>1.032</td>
<td>1.032</td>
<td>1.043</td>
<td>1.034</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>1.421</td>
<td>1.421</td>
<td>1.421</td>
<td>1.421</td>
<td>1.421</td>
<td>1.421</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>1.872</td>
<td>1.905</td>
<td>1.907</td>
<td>1.907</td>
<td>1.799</td>
<td>1.755</td>
</tr>
<tr>
<td>3</td>
<td>712</td>
<td>Low limit</td>
<td>0.575</td>
<td>0.576</td>
<td>0.576</td>
<td>0.575</td>
<td>0.572</td>
<td>0.557</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>1.074</td>
<td>1.095</td>
<td>1.096</td>
<td>1.096</td>
<td>1.038</td>
<td>1.033</td>
</tr>
<tr>
<td>4</td>
<td>209</td>
<td>Low limit</td>
<td>0.799</td>
<td>0.802</td>
<td>0.801</td>
<td>0.800</td>
<td>0.789</td>
<td>0.794</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>1.351</td>
<td>1.351</td>
<td>1.351</td>
<td>1.351</td>
<td>1.351</td>
<td>1.351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>2.048</td>
<td>2.126</td>
<td>2.136</td>
<td>2.136</td>
<td>1.913</td>
<td>1.898</td>
</tr>
<tr>
<td>5</td>
<td>539</td>
<td>Low limit</td>
<td>0.424</td>
<td>0.425</td>
<td>0.425</td>
<td>0.425</td>
<td>0.386</td>
<td>0.374</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>0.744</td>
<td>0.744</td>
<td>0.744</td>
<td>0.744</td>
<td>0.744</td>
<td>0.744</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>1.153</td>
<td>1.201</td>
<td>1.208</td>
<td>1.208</td>
<td>1.102</td>
<td>1.090</td>
</tr>
<tr>
<td>6</td>
<td>1206</td>
<td>Low limit</td>
<td>0.475</td>
<td>0.476</td>
<td>0.475</td>
<td>0.475</td>
<td>0.471</td>
<td>0.475</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>0.651</td>
<td>0.651</td>
<td>0.651</td>
<td>0.651</td>
<td>0.651</td>
<td>0.651</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>0.856</td>
<td>0.871</td>
<td>0.872</td>
<td>0.872</td>
<td>0.832</td>
<td>0.824</td>
</tr>
<tr>
<td>7</td>
<td>386</td>
<td>Low limit</td>
<td>0.689</td>
<td>0.690</td>
<td>0.690</td>
<td>0.690</td>
<td>0.692</td>
<td>0.693</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SMR</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
<td>0.990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper limit</td>
<td>1.345</td>
<td>1.375</td>
<td>1.377</td>
<td>1.377</td>
<td>1.288</td>
<td>1.267</td>
</tr>
</tbody>
</table>
Table 6.2. risk distributions of patients operated on by 7 surgeons

<table>
<thead>
<tr>
<th>Parsonnet score</th>
<th>Low</th>
<th>Elevated</th>
<th>Significantly elevated</th>
<th>High</th>
<th>Very high</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>29.6%</td>
<td>22.1%</td>
<td>14.7%</td>
<td>15.3%</td>
<td>18.3%</td>
</tr>
<tr>
<td>2</td>
<td>19.7%</td>
<td>23.7%</td>
<td>22.7%</td>
<td>13.3%</td>
<td>20.6%</td>
</tr>
<tr>
<td>3</td>
<td>34.7%</td>
<td>26.1%</td>
<td>15.2%</td>
<td>12.5%</td>
<td>11.5%</td>
</tr>
<tr>
<td>4</td>
<td>34.4%</td>
<td>21.5%</td>
<td>15.8%</td>
<td>14.4%</td>
<td>13.9%</td>
</tr>
<tr>
<td>5</td>
<td>54.7%</td>
<td>28.2%</td>
<td>12.2%</td>
<td>3.3%</td>
<td>1.5%</td>
</tr>
<tr>
<td>6</td>
<td>41.5%</td>
<td>25.0%</td>
<td>15.0%</td>
<td>9.5%</td>
<td>9.0%</td>
</tr>
<tr>
<td>7</td>
<td>20.2%</td>
<td>23.3%</td>
<td>17.6%</td>
<td>14.8%</td>
<td>24.1%</td>
</tr>
<tr>
<td>Phase I</td>
<td>41.7%</td>
<td>23.3%</td>
<td>13.9%</td>
<td>9.0%</td>
<td>12.1%</td>
</tr>
<tr>
<td>Phase II</td>
<td>35.3%</td>
<td>24.4%</td>
<td>15.4%</td>
<td>11.7%</td>
<td>13.2%</td>
</tr>
</tbody>
</table>

§7. Conclusion

Traditionally, confidence intervals for SMR are usually constructed by assuming (i) the numerator $O$ is distributed as a Poisson random variable, and (ii) the denominator of the SMR, $\sum_{i=1}^{n} X_i$, is a constant. The Poisson assumption may not be valid for some applications. In this thesis, we do not make these two assumptions, instead we derive the asymptotic distribution of the SMR and use it to construct a confidence interval for the SMR. We also develop another confidence interval based on the bootstrap distribution of SMR. Our study shows that the estimated coverage probabilities of the confidence intervals based on the asymptotic distribution and bootstrap approach are close to specified level. The estimated coverage probabilities of the four confidence intervals based on traditional approach are consistently above the specified level. In other words, these four confidence intervals are too conservative. The Byar’s and Ulm’s approaches are conservative because they are developed using the Clopper-Pearson approach. The Vandenbroucke’s and Ury’s approaches are conservative because the variance of the SMR is larger as a result of the two assumptions. In terms of power, our study shows that the power of the
6 confidence intervals are approximately the same when the odds ratio of death $Q$ is greater than the odds ratio of death under the null hypothesis, $Q_0$. However, the power of the asymptotic and bootstrap approaches are significantly higher than the other approaches when $Q < Q_0$. The difference in power is due to a difference between the normal and Poisson approximations. The Poisson approximation is right-skewed with a larger variance. This explains why the upper confidence limits of the SMR based on the Poisson approximation is larger and resulted in lower power because rejection is based on the upper confidence limit when $Q < Q_0$. The effects of the risk distributions of patients on SMR were also studied. The results show that a worse than average surgeon can “improve” his SMR by operating on more high risk patients. Similarly, a better than average surgeon can “improve” his performance in SMR by operating on more low risk patients. The latter is contrary to a popular believe that a surgeon can only “improve” his performance by operating on low risk patients (Kahn et al., 2007). It should be noted that a significant change in SMR can only be effected if there are substantial changes in a risk distribution. In conclusion, we have provided two new confidence intervals for the SMR which are more accurate than the traditional ones. Finally, note that a large sample is usually required for estimating the SMR because the death rate is usually very low. If the number of deaths in a sample is zero, then no confidence interval can be constructed. For a small sample with a small number of deaths, the length of the confidence interval might not be small enough for it to be useful.
Appendix A: R Codes for Calculating Coverage Probabilities

```
sgh=read.table("SGH1_data.txt");
record=sgh[,13];parsonnet=sgh[,15];
nrecord=length(record);
record1=record2=rep(0,nrecord);
for(i in 1:nrecord)
  {if(record[i]<31&record[i]>-1) record1[i]=1 else record2[i]=1}
record3=cbind(record1,record2);
logitm=glm(record3~parsonnet,family=binomial("logit"));
logp=predict(logitm,data.frame(parsonnet=parsonnet));
po1=parsonnet;
po2=read.table("po2.txt");po2=po2[,1];
po3=read.table("po3.txt");po3=po3[,1];

################get the true risk of patients#################
getrisk=function(parsonnetscore,Odds){
  logitp=predict(logitm,data.frame(parsonnet=parsonnetscore));
  risk=exp(logitp)/(1+exp(logitp));
  truerisk=Odds*risk/(1-risk+Odds*risk);
  return(truerisk);
}

################get outcome for patients####################
getoutcome=function(parsonnetscore,Odds){
  truerisk=getrisk(parsonnetscore,Odds);
  number=length(parsonnetscore);
  test=runif(number);
  result=rep(0,number);
  for(i in 1:number) {if(test[i]<truerisk[i]) result[i]=1}
  return(result);
}

########################################################################
Getsample=function(samplesize, odds){
  Pars=sample(po1,samplesize,TRUE);
  Outs=getoutcome(Pars,odds);
```
Datas = cbind(Pars, Outs);
return(Datas);

########## World wide data

Dataw = read.table("Dataw_low.txt")
Parw = Dataw[,1];
Outw = Dataw[,2];
Modelw = glm(Outw ~ Parw, family = binomial("logit"));

Vanci = function(par, out, alpha) {
o = sum(out);
risk = predict(Modelw, data.frame(Parw = par));
risk = exp(risk) / (1 + exp(risk));
e = sum(risk);
\[ c_i = \frac{\left( \sqrt{o} - \frac{qnorm(1 - alpha/2)}{2} \right)^2, \left( \sqrt{o} + \frac{qnorm(1 - alpha/2)}{2} \right)^2}{e}; \]
return(ci);
}

Vancheck = function(par, out, alpha) {
  ci = Vanci(par, out, alpha);
  if (ci[1] < 1 & ci[2] > 1) mark = 1 else mark = 0
  return(mark);
}

Uryci = function(par, out, alpha) {
o = sum(out);
risk = predict(Modelw, data.frame(Parw = par));
risk = exp(risk) / (1 + exp(risk));
e = sum(risk);

if (alpha == 0.05)
ci=c(o-qnorm(1-alpha/2)*sqrt(o)+1, o+qnorm(1-alpha/2)*sqrt(o)+2)/e
else
   ci=c(o-qnorm(1-alpha/2)*sqrt(o)+2, o+qnorm(1-alpha/2)*sqrt(o)+3)/e
}
return(ci);
}

Urycheck=function(par,out,alpha){
  ci=Uryci(par,out,alpha);
  if(ci[1]<1&ci[2]>1) mark=1 else mark=0
  return(mark);
}

Ulmci=function(par,out,alpha){
o=sum(out);
  risk=predict(Modelw,data.frame(Parw=par));
  risk=exp(risk)/(1+exp(risk));
  e=sum(risk);
  l=qchisq(alpha/2,2*o)/2;
  u=qchisq(1-alpha/2,2*(o+1))/2;
  ci=c(l,u)/e;
  return(ci);
}

Ulmcheck=function(par,out,alpha){
  ci=Ulmci(par,out,alpha);
  if(ci[1]<1&ci[2]>1) mark=1 else mark=0
  return(mark);
}

Byarci=function(par,out,alpha){
o=sum(out);
```r
risk=predict(Modelw,data.frame(Parw=par));
risk=exp(risk)/(1+exp(risk));
e=sum(risk);
l=(1-1/(9*o)-qnorm(1-alpha/2)/(3*sqrt(o)))^3*o/e;
u=(1-1/(9*(o+1))+qnorm(1-alpha/2)/(3*sqrt(o+1)))^3*(o+1)/e;
 ci=c(l,u);
return(ci);
}

Byarcheck=function(par,out,alpha){
ci=Byarci(par,out,alpha);
if(ci[1]<1&ci[2]>1) mark=1 else mark=0
return(mark);
}

.asci=function(par,out,alpha){
risk=predict(Modelw,data.frame(Parw=par));
risk=exp(risk)/(1+exp(risk));
mux=mean(risk);
muy=mean(out);
sigx=var(risk);
sigy=var(out);
sigxy=cov(risk,out);
smr=muy/mux;
vs=(muy^2*sigx/mux^4-2*muy*sigxy/mux^3+sigy/mux^2)/length(out);
sd=sqrt(vs);
 ci=c(smr-qnorm(1-alpha/2)*sd, smr+qnorm(1-alpha/2)*sd);
return(ci);
}

ascheck=function(par,out,alpha){
ci=asci(par,out,alpha);
if(ci[1]<1&ci[2]>1) mark=1 else mark=0
return(mark);
}
```

alsci=function(par,out,alpha){
  risk=predict(Modelw,data.frame(Parw=par));
  risk=exp(risk)/(1+exp(risk));
  mux=mean(risk);
  muy=mean(out);
  sigx=var(risk);
  sigy=var(out);
  sigxy=cov(risk,out);
  lsmr=log(muy/mux);
  vs=(sigx/mux^2-2*sigxy/(mux*muy)+sigy/muy^2)/length(out);
  sd=sqrt(vs);
  ci=c(lsmr-qnorm(1-alpha/2)*sd, lsmr+qnorm(1-alpha/2)*sd);
  return(ci);
}

alscheck=function(par,out,alpha){
  ci=alsci(par,out,alpha);
  if(ci[1]<0&ci[2]>0) mark=1 else mark=0
  return(mark);
}

csmr=function(par,out){
  o=sum(out);
  risk=predict(Modelw,data.frame(Parw=par));
  risk=exp(risk)/(1+exp(risk));
  e=sum(risk);
  smr=o/e;
  return(smr);
}

bsci=function(par,out,alpha1,alpha2,numberofbootstrap){
  smr=csmr(par,out);
}
Risk-Adjusted CUSUM Charts and SMR

Tang Xu

```r
record = rep(0, numberofbootstrap);
samplesize = length(par);
index = seq(1, samplesize);
for(i in 1:numberofbootstrap){
    indext = sample(index, samplesize, TRUE);
    part = par[indext];
    outt = out[indext];
    record[i] = csmr(part, outt);
}
q1 = c(alpha1/2, 1-alpha1/2);
quantiles1 = quantile(record, q1);
l1 = 2*smr - quantiles1[2];
u1 = 2*smr - quantiles1[1];
ci1 = c(l1, u1);
q2 = c(alpha2/2, 1-alpha2/2);
quantiles2 = quantile(record, q2);
l2 = 2*smr - quantiles2[2];
u2 = 2*smr - quantiles2[1];
ci2 = c(l2, u2);
ci = c(ci1, ci2);
return(ci);
}
bscheck = function(par, out, alpha1, alpha2, numberofbootstrap){
ci = bsci(par, out, alpha1, alpha2, numberofbootstrap);
if(ci[1]<1 & ci[2]>1) mark1 = 1 else mark1 = 0
mark = c(mark1, mark2);
return(mark);
}
clsmr = function(par, out){
```

127
o=sum(out);
risk=predict(Modelw,data.frame(Parw=par));
risk=exp(risk)/(1+exp(risk));
e=sum(risk);
lsmr=log(o/e);
return(lsmr);
}

blsci=function(par,out,alpha1,alpha2,numberofbootstrap){
  lsmr=clsmr(par,out);
  record=rep(0,numberofbootstrap);
samplesize=length(par);
index=seq(1,samplesize);
for(i in 1:numberofbootstrap){
  indext=sample(index,samplesize, TRUE);
  part=par[indext];
  outt=out[indext];
  record[i]=clsmr(part,outt);
}

q1=c(alpha1/2,1-alpha1/2);
quantiles1=quantile(record,q1);
l1=2*lsmr-quantiles1[2];
u1=2*lsmr-quantiles1[1];
ci1=c(l1,u1);

q2=c(alpha2/2,1-alpha2/2);
quantiles2=quantile(record,q2);
l2=2*lsmr-quantiles2[2];
u2=2*lsmr-quantiles2[1];
ci2=c(l2,u2);

return(ci);
blscheck=function(par,out,\alpha_1,\alpha_2,\text{numberofbootstrap}){
  ci=blsci(par,out,\alpha_1,\alpha_2,\text{numberofbootstrap});
  if(ci[1]<0\&ci[2]>0) mark_1=1 else mark_1=0
  if(ci[3]<0\&ci[4]>0) mark_2=1 else mark_2=0
  mark=c(mark_1,mark_2);
  return(mark);
}

#########################################################################
coverage=function(odds,\alpha_1,\alpha_2,\text{numberofbootstrap},\text{timesoftrials}){
  Vanrecord=Uryrecord=Ulmrecord=Byarrecord=asrecord=
  alsrecord=bsrecord=blsrecord=matrix(0,\text{timesoftrials},2);
  for(i in 1:\text{timesoftrials}){
    sample=Getsample(1000,odds);
    pars=sample[,1];
    outs=sample[,2];
    Vanrecord[i,]=c(Vancheck(pars,outs,\alpha_1), Vancheck(pars,outs,\alpha_2));
    Uryrecord[i,]=c(Urycheck(pars,outs,\alpha_1), Urycheck(pars,outs,\alpha_2));
    Ulmrecord[i,]=c(Ulmcheck(pars,outs,\alpha_1), Ulmcheck(pars,outs,\alpha_2));
    Byarrecord[i,]=c(Byarcheck(pars,outs,\alpha_1), Byarcheck(pars,outs,\alpha_2));
    asrecord[i,]=c(ascheck(pars,outs,\alpha_1),ascheck(pars,outs,\alpha_2));
    alsrecord[i,]=c(alscheck(pars,outs,\alpha_1),alscheck(pars,outs,\alpha_2));
    bsrecord[i,]=bscheck(pars,outs,\alpha_1,\alpha_2,\text{numberofbootstrap});
    blsrecord[i,]=blscheck(pars,outs,\alpha_1,\alpha_2,\text{numberofbootstrap});
  }
  Van=c(mean(Vanrecord[,1]),mean(Vanrecord[,2]));
  Ury=c(mean(Uryrecord[,1]),mean(Uryrecord[,2]));
  Ulm=c(mean(Ulmrecord[,1]),mean(Ulmrecord[,2]));
  Byar=c(mean(Byarrecord[,1]),mean(Byarrecord[,2]));
  as=c(mean(asrecord[,1]),mean(asrecord[,2]));
  als=c(mean(alsrecord[,1]),mean(alsrecord[,2]));
  bs=c(mean(bsrecord[,1]),mean(bsrecord[,2]));
  bls=c(mean(blsrecord[,1]),mean(blsrecord[,2]));
result=rbind(Van,Ury,Ulm,Byar,as,als,bs,bls); return(result);
References


Canadian Institute for Health Information (2012), *Technical Notes: Hospital Standardized Mortality Ratio (HSMR)*. Ottawa, Canada: The Institute.

Clopper, C. J. and Pearson, E. S. (1934), “The Use of Confidence or Fiducial limits Illustrated in the Case of the Binomial.” *Biometrika*, 26, 404–413.


Chapter 3. Comparison of Standardized Mortality Ratios

Abstract

Measuring quality of medical practice is becoming increasingly prominent in quality management because it is a key component in improving quality and efficiency in health care. The traditional standardized mortality ratio (SMR) compares the mortality rate of a study population with that of a reference population by comparing the actual number of deaths in the study population with the expected number of deaths in the study population assuming the mortality rate of the reference population. In order to measure the performance of a surgeon or a group of surgeons in a hospital in performing a particular type of surgical operation, for example the heart bypass operation, the SMR is used to compare the observed number of deaths in a sample with the estimated number of deaths usually calculated based on the average performance of a group of surgeons in a hospital, region, country or certain parts of the world. The estimated number of deaths in a sample associated with this new application of SMR is not a constant but a random variable. This means that all existing results for the traditional SMR may no longer be valid for the new SMR. It seems to make sense to use a regional, national or some world-wide data set as a reference data set to estimate the number of deaths and to compare this number with the observed number of deaths. However, we discover that comparison of SMRs may not be meaningful because of the different risk distributions or distributions of patients, despite the fact that the SMR has taken the risk distribution of patients into account. In this thesis, we develop two procedures for comparing SMRs. The asymptotic distribution of the test statistics are derived. We also use the bootstrap procedure to estimate the finite-sample distributions. Both type I errors and powers of these procedures are investigated. Empirical evidence shows that our procedures are more sensitive than an existing method developed for the traditional SMR but adapted for the new SMR.

KEY WORDS: Bootstrap; Cardiac operations; Euroscores; Hospital standardized mortality ratio; Logistic regression; Odds ratio; Parsonnet scores; Patient mix; Quality monitoring; risk distributions.
§1. Introduction

The traditional standardized mortality ratio (SMR) is defined as $O/E$ where $O$ is the actual number of deaths in a study population and $E$ is the expected number of deaths in the study population assuming the mortality rate of a reference population. Hence, the SMR allows the mortality rate in a study population to be compared with that of a reference population. According to Keiding (1987), Dale (1777) was probably the earliest to explain and use the SMR. The SMR is described in many biostatistics textbooks, for example Forthofer et al. (2007, page 55), Armitage et al. (2008, pages 659–666) and Rosner (2011, pages 253–256). It is popularly used in various fields including occupational epidemiology (Ulm, 1989; Reid et al., 2008; Mirabelli et al., 2008), cancer studies (Breslow and Day, 1987; Mastrangelo et al., 2008; Gun et al., 2008; Guha, 2010; Bagary, 2011), heart studies (Humblet et al., 2008; Hickey et al., 2011; Nieuwkamp et al., 2011) and medical care (Thomas and Hofer, 1999; Sonnenberg, 2008; Zeger et al., 2008; Gensberg et al., 2009; Steenland et al., 2010; Beard et al., 2011).

Quite often in practice, instead of an entire study population, a sample is taken from the study population and the statistic $O$ now represents the observed number of deaths in the sample. The quantity $E$ represents the expected number of deaths in the sample assuming the mortality rate of a reference population. In order to derive a confidence interval for the SMR, the random variable $O$ is usually assumed to follow a Poisson distribution (Vandenbroucke, 1982; Ury, 1985 and Ulm, 1989). The Poisson assumption is justified in cases where the deaths can be viewed as occurrences of rare events. Approximate confidence intervals for the SMR are obtained by first deriving a confidence interval for the expectation of $O$, and then dividing the interval with $E$. Vandenbroucke (1982) approximated the square root of a Poisson random variable with a normal random variable to obtain a 95% confidence interval for $O$ as $([\sqrt{O} - 1]^2, [\sqrt{O} + 1]^2)$. Ury (1985) proposed $(O - 1.96\sqrt{O} + 1, O + 1.96\sqrt{O} + 2)$ as an improvement. Ulm (1989) used
the relationship between the Poisson and chi-square probabilities \( \sum_{i=0}^{m-1} e^{-\lambda} \frac{\lambda^i}{i!} = \Pr(\chi^2_{2m} > 2\lambda) \), to obtain 100(1 - \alpha)% confidence limits \( \lambda_L \) and \( \lambda_U \) based on \( O \) where \( \Pr(\chi^2_{2O} \leq 2\lambda_L) = \alpha/2 \) and \( \Pr(\chi^2_{2(O+1)} \leq 2\lambda_U) = 1 - \alpha/2 \). A widely used confidence interval due to Byar (Breslow and Day, 1987) is given as

\[
\left( \left[ 1 - \frac{1}{9O} - \frac{z_{\alpha/2}}{3\sqrt{O}} \right]^3 \frac{O}{E}, \left[ 1 - \frac{1}{9(O+1)} + \frac{z_{\alpha/2}}{3\sqrt{O+1}} \right]^3 \frac{O+1}{E} \right),
\]

where \( z_{\alpha/2} \) is the 100(1 - \alpha/2)th standard normal percentile. It should be noted that the Poisson assumption may not always be true. In addition, for cases where there is a specific mortality rate for a population, the statistic \( O \) is a binomial random variable because the outcomes for the individuals are independent and identically distributed Bernoulli random variables, and the confidence interval derived using the binomial distribution will be exact.

If there are \( k \) different mortality rates for \( k \) groups of a study population defined by age for example (see Kielding, 1987), the SMR can be written as \( \text{SMR} = \frac{O}{E} = \sum_{i=1}^{k} \frac{N_i\alpha_i}{\sum N_i\lambda_i} \) where \( N_i \) is subpopulation size of the \( i \)th group in the study population, \( \alpha_i \) and \( \lambda_i \) are the mortality rates of the \( i \)th group of the study and reference populations respectively. As before, the numerator \( O \) is the actual number of deaths in the study population and the denominator \( E \) is the expected number of deaths in the study population assuming the mortality rates of the reference population. If a random sample is taken from the study population, then the resulting SMR is given by \( \text{SMR} = \frac{O}{\sum n_i\lambda_i} \) where \( O \) is the observed number of deaths in the sample and \( n_i \) is the subsample size of the \( i \)th group in the sample. Note that the denominator is no longer a constant but a random variable determined by the subsample sizes in a given sample. It can be shown that \( E(\sum n_i\lambda_i) = \frac{n}{N} \sum N_i\lambda_i \)

where \( n = n_1 + \ldots + n_k \) and \( N = N_1 + \ldots + N_k \). This shows that \( \sum n_i\lambda_i \) is an estimate of the expected number of deaths in a sample of size \( n \).

Measuring quality of medical practice is becoming increasingly prominent in quality management because it is a key component in improving quality and effi-
ciency in health care. The hospital standardized mortality ratio (HSMR) was developed by Jarman et al. (1999) for measuring performances of hospitals. According to the HSMR Technical Notes published by the Canadian Institute for Health Information (2012), the HSMR is defined as the observed number of deaths in a hospital divided by the expected number of deaths based on 65 diagnostic groups which account for 80% of in-hospital deaths, excluding patients who received palliative care. If the risk distributions of patients remain the same, then the HSMR tracked over time indicates how successful hospitals and health regions have been in reducing in-hospital deaths, and this could lead to improved medical care. In the USA, the Institute of Health Improvement (Whittington, 2005) is using the HSMR in their campaigns to improve the safety of patients by implementing strategies to reduce mortality. In England, the National Patient Safety Agency (Thomson, 2007) has adopted the HSMR as a high level track measure for patient safety. The Canadian Institute for Health Information has led the effort in promoting the use of HSMR for Canada and publishes results for eligible facilities and regions in all provinces outside Quebec. Publications of HSMRs of hospitals will inevitably lead to direct and often naive comparisons and rankings of hospitals. As we shall demonstrate in Section 2 that such comparisons and rankings might not be meaningful because of differences in risk distributions of patients, despite the fact that the HSMR has taken risk distributions into account.

Although the term ‘expected’ is used in the definition of HSMR, it is clear from the way that the ‘expected’ number of deaths is defined and calculated according to Appendix I of the HSMR Technical Notes published by the Canadian Institute for Health Information (2012), it is an ‘estimated’ number of deaths because it depends on the risk distributions of patients and changes from one sample to another. In other words, the estimated number of deaths is a random variable and clearly cannot be an expected quantity. Thus, existing results derived for the traditional SMR may no longer be valid for the HSMR. For example, the Byar’s confidence interval was
derived for the traditional SMR and it is now widely used for the HSMR (see for examples, Guidelines for Using and Developing Rates for Public Health Assessment, 2002 and HSMR Technical Notes published by the Canadian Institute for Health Information, 2012). In order to obtain a confidence interval for HSMR using the Byar’s confidence interval, the expected number of deaths $E$ in Byar’s confidence interval is simply replaced by an estimate of the expected number of deaths. This casts doubts on the validity of the resulting confidence interval. Faris et al. (2003) showed using simulation that such an approach will result in a wider confidence interval for the SMR.

When the SMR is used in this new application, we obtain a new SMR because we are comparing an observed number of deaths with an estimate of the expected number of deaths. The denominator of the new SMR is not a constant but a random variable. All existing theoretical results for the traditional SMR may no longer be valid for the new SMR unless the estimate is close to the expected number of deaths. If a sample is viewed as a population, then it is not meaningful to study the variance of SMR because the SMR is now assumed to be a constant. In Section 2, we formally introduce the new SMR and derive its statistical properties. The effect of risk distribution on SMR will be studied. The asymptotic distribution of the SMR is also derived. In Sections 3 and 4, we develop procedures for comparing two SMRs but the finite-sample distributions of the test statistics are mathematically intractable. We use the bootstrap procedure to estimate the finite-sample distributions for constructing confidence intervals. Both type I errors and powers of these procedures are investigated. Examples are given in Section 5 and a conclusion is given in Section 6.

§2. **Statistical Properties of SMR**

2.1 **Definition and Expectation of SMR**

Suppose a patient is to undergo a cardiac surgery. The surgical outcome of the patient can be represented by $Y$ which is 1 if the patient dies within 30 days and
0 otherwise. The patient’s conditions like age, blood pressure, existence of certain
disease like diabetes, morbid obesity et al. will be determined and this information
can be summarized as a Parsonnet score (Parsonnet et al., 1989). The Parsonnet
score is an integer from 0 to 100 and it is a measure of the risk of death of a patient
who undergoes a cardiac surgery. A patient’s Parsonnet score $S$ can then be used
to estimate the probability of death $X$ by using a logistic regression model $m(\cdot)$
built using some past surgical data as $X = m(S)$. The quantity $X$ is an estimate
of the probability of death assuming the average performance of all surgeons in the
data. The model defines a one-to-one relationship between $S$ and $X$. For a patient
who is operated on by a particular surgeon, let $d(X)$ be the true risk of death
and define $D(S) = d(m(S))$. This means that given a patient with Parsonnet
score $S$, the true risk of death operated on by the surgeon will be represented by
$D(S)$. If a model is fitted using a different data set, $m(\cdot)$ and $d(\cdot)$ will change but
$D(\cdot) = d(m(\cdot))$ will remain the same. In effect, we have defined $d(\cdot)$ and $D(\cdot)$ as
measures of performance of a surgeon. A surgeon who performs uniformly better
than a surgeon with an average performance will have $D(s) < m(s)$ for all $s$ or
$d(x) < x$ for all $x$. Similarly, a surgeon who performs uniformly worse than average
will have $D(s) > m(s)$ for all $s$ or $d(x) > x$ for all $x$.

Let $S_1, S_2, \cdots, S_n$ be the Parsonnet scores of a random sample of $n$ patients.
The new SMR is defined as

$$\text{SMR} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}, \quad (1)$$

where $\sum_{i=1}^{n} Y_i$ is the observed number of deaths and $\sum_{i=1}^{n} X_i$ is an estimate of the
expected number of deaths in a sample of size $n$ given the model $x = m(s)$. The
expected number of deaths in the sample of size $n$ is calculated based on the average
performance of all the surgeons whose data were used to fit the model $x = m(s)$.
A model $x = m(s)$ that underestimates the risk $x$ will inflate the SMR. Similarly,
a model that overestimates the risk $x$ will deflate the SMR. The SMR is thus
dependent on how the risk $x$ is estimated and hence it is subjected to possible abuses or manipulations.

Given the estimated risk $X = x$ of a patient, $Y$ is a Bernoulli random variable with parameter $d(x)$. The SMR is a ratio of two random variables and using Taylor’s approximation, it can be shown that

$$E(\text{SMR}) \approx \frac{\mu_Y}{\mu_X}, \quad (2)$$

where $\mu_X = \sum xp(x)$, $\mu_Y = \sum d(x)p(x)$ and $p(\cdot)$ is the probability mass function (pmf) of $X$. Thus, the SMR is an almost unbiased estimator of $\frac{\mu_Y}{\mu_X}$. The quantity $\frac{\mu_Y}{\mu_X}$ can also be expressed as

$$\frac{\mu_Y}{\mu_X} = \sum \frac{d(x)}{x} \cdot \frac{xp(x)}{\mu_X}, \quad (3)$$

which can be viewed as a weighted average of the performance measure $d(x)/x$ with weight $xp(x)/\mu_X$. The weight depends on both $x$ and $p(x)$, hence $\frac{\mu_Y}{\mu_X}$ is influenced by both the estimated risk of death and its distribution. Equation (3) also indicates that for two equally competent surgeons with the same $d(\cdot)$, their SMRs can be different, depending on the risk distribution of their patients. Therefore, a direct comparison of the magnitudes of SMRs may not be meaningful when their patients’ risk distributions are different.
Figure 2.1. Plots of true probability of death $d(x)$ of a patient operated on by a surgeon characterized by the odds ratio of death $Q$ against the probability of death $x$ of a patient operated on by a surgeon with an average performance.
2.2 Effect of Risk Distribution on SMR

In this section, we will examine the effect of risk distribution of patients on the SMR. Suppose the odds ratio of death associated with a surgeon is $Q = [d(x)/(1 - d(x))]/[x/(1 - x)]$ (Steiner et al. 2000). This means the odds of death of a patient operated on by the surgeon is $Q$ times the odds of death if the patient were to be operated on by a surgeon of average performance. A plot of $d(x)$ against $x$ for $Q = 0.5, 1$ and 2 is displayed in Figure 2.1. For a surgeon with $Q = 1$, the true risk of death is represented by the diagonal line $d(x) = x$. Note that the true risk of death of a patient operated on by a surgeon with $Q > 1$ is larger than that of $Q = 1$. The opposite is true for $Q < 1$.

![Frequency distribution of Parsonnet scores](image)

*Figure 2.2.* The frequency distribution of Parsonnet scores of a group of 6449 patients who underwent cardiac operations.
Consider a real data set of 6449 Parsonnet scores of cardiac patients and its histogram is shown in Figure 2.2. Each patient was operated on by a surgeon from a group of surgeons. This data set is denoted as population $P_0$. A real data set was used for the study so that any results obtained will be representative of real scenarios. A logistic regression model

$$
\log(x_0/(1 - x_0)) = \text{logit}(x_0) = \alpha_0 + \beta_0 s,
$$

is first fitted using $P_0$ and its associated surgical outcomes to obtain a relationship between the Parsonnet score $s$ and the risk of death $x$ assuming the average performance of all the surgeons. The two parameters were found to be $\alpha_0 = -3.605$ and $\beta_0 = 0.072$.

For the study, we take random samples of Parsonnet scores with replacement from $P_0$ to study the effect of risk distribution of patients on SMR. Samples of 100 Parsonnet scores are taken randomly from each of the five risk categories of $P_0$ as classified in National Adult Cardiac Surgical Database Report 2000-2001: low (0–4), elevated (5–9), significantly elevated (10–14), high (15–19), very high (20–100). The risk of death of a patient assuming the performance of an average surgeon, $x_0$ (and represented as $x$) is calculated using equation (4) and $x$ is assumed to be the true risk of death in the study. For each of the 5 samples, the patients are operated on by a surgeon with odds ratio of death $Q = 0.5$. In other words, the true risk of death of a patient operated on by the surgeon is given as $d(x) = Qx/(Qx + 1 - x)$. Since the pmf of $X$ is known for $P_0$, the quantity $\mu_Y/\mu_X$ can be calculated exactly using equation (2). This is repeated for a surgeon with $Q = 0.6, 0.7, \ldots, 2.0$.

Table 2.1 contains the quantity $\mu_Y/\mu_X$ classified by the risk category of patients and the odds ratio of death associated with a surgeon. The table reveals that the SMR of a surgeon with an average performance ($Q = 1$) is unaffected by the risk distribution. This is due to the fact that if $Q = 1$, then $d(x) = x$ and equation (2) yields $\mu_Y/\mu_X = 1$. For a surgeon who performs better than average ($Q < 1$), the SMR is larger if the risks are of a higher category. In contrast, for a surgeon who
performs worse than average ($Q > 1$), the SMR is smaller if the risks are of a higher category. This is contrary to a popular misconception among hospital administrators and medical professionals that even risk-adjusted performance measures will fare worse if the patients operated on are of higher risks. For example, Kahn et al. (2007) commented that hospitals might shift the obligations of treating higher risk patients by transferring them out to other health care institutions in order to get a smaller SMR. The table shows that this is not true unless $Q$ is less than 1. For $Q$ closer to 1, the quantity $\mu_Y/\mu_X$ varies less across the risk categories. This shows that the SMR does in fact adjust for the risks of patients. However, for a surgeon with $Q = 0.5$ for example, the SMR for the very high risk category can be 20% larger than that for a low risk category. Also, the SMR of a surgeon with $Q = 2.0$ operating on very high risk patients is smaller than the SMR of a surgeon with $Q = 1.8$ operating on low risk patients. Thus, it may not be meaningful to compare the magnitudes of two SMRs directly if their risk distributions are different.

Table 2.1. True SMRs of 11 surgeons operated on patients from 5 different risk groups

<table>
<thead>
<tr>
<th>Risk category</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
<th>1.2</th>
<th>1.4</th>
<th>1.6</th>
<th>1.8</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.51</td>
<td>0.61</td>
<td>0.71</td>
<td>0.80</td>
<td>0.90</td>
<td>1.19</td>
<td>1.38</td>
<td>1.57</td>
<td>1.76</td>
<td>1.94</td>
<td></td>
</tr>
<tr>
<td>Elevated</td>
<td>0.51</td>
<td>0.61</td>
<td>0.71</td>
<td>0.81</td>
<td>0.90</td>
<td>1.19</td>
<td>1.38</td>
<td>1.56</td>
<td>1.74</td>
<td>1.92</td>
<td></td>
</tr>
<tr>
<td>Significantly elevated</td>
<td>0.52</td>
<td>0.61</td>
<td>0.71</td>
<td>0.81</td>
<td>0.91</td>
<td>1.19</td>
<td>1.37</td>
<td>1.54</td>
<td>1.72</td>
<td>1.89</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.52</td>
<td>0.62</td>
<td>0.72</td>
<td>0.81</td>
<td>0.91</td>
<td>1.18</td>
<td>1.35</td>
<td>1.52</td>
<td>1.68</td>
<td>1.84</td>
<td></td>
</tr>
<tr>
<td>Very high</td>
<td>0.60</td>
<td>0.69</td>
<td>0.78</td>
<td>0.86</td>
<td>0.93</td>
<td>1.13</td>
<td>1.25</td>
<td>1.36</td>
<td>1.46</td>
<td>1.55</td>
<td></td>
</tr>
</tbody>
</table>

2.3 Asymptotic Distribution of SMR and log(SMR)

We will derive the asymptotic distribution of SMR and log(SMR). The reason for studying log(SMR) is that even though the asymptotic distribution of SMR is normal but the finite-sample distribution is found to be substantially right-skewed
even when the sample size is very large. As a result, the power of test procedures based on SMR is found to be affected significantly by the skewness.

**Theorem 1.** Consider a random sample of \( n \) patients with their probabilities of death and surgical outcomes denoted as \((X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\), then the SMR = \( \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} \) has the following asymptotic distribution

\[
\sqrt{n}(\text{SMR} - \frac{\mu_Y}{\mu_X}) \xrightarrow{L} N(0, \frac{\mu_Y^2}{\mu_X^2} \sigma_X^2 - 2 \frac{\mu_Y}{\mu_X} \sigma_{XY} Y + \frac{1}{\mu_X^2} \sigma_Y^2),
\]

where \( \mu_X = E(X), \mu_Y = E(Y), \sigma_X^2 = \text{Var}(X), \sigma_Y^2 = \text{Var}(Y) \) and \( \sigma_{XY} = \text{Cov}(X, Y) \).

**Proof.** Let \( W_i = (X_i, Y_i)^T \) and note that \( W_1, W_2, \ldots, W_n \) are independent and identically distributed (iid) random vectors. The expectation and covariance matrix of \( W_i \) are given as \( E(W_i) = (\mu_X, \mu_Y)^T \) and

\[
\text{Var}(W_i) = \begin{pmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{pmatrix}.
\]

Let \( f(x, y) = y/x \), then SMR = \( f(\bar{X}, \bar{Y}) \). Also, define \( \nabla^T f(\mu_X, \mu_Y) = (-\mu_Y/\mu_X^2, 1/\mu_X) \).

Using Delta method, it can be shown that

\[
\sqrt{n}(f(\bar{X}, \bar{Y}) - f(\mu_X, \mu_Y)) \xrightarrow{L} N(0, \tau^2),
\]

where \( \tau^2 = \nabla^T f(\mu_X, \mu_Y) \text{Var}(W_i) \nabla f(\mu_X, \mu_Y) = \frac{\mu_Y^2}{\mu_X^4} \sigma_X^2 - 2 \frac{\mu_Y}{\mu_X^2} \sigma_{XY} + \frac{1}{\mu_X^2} \sigma_Y^2 \). This completes the proof.

Note that for this theorem, the risk of death of a patient \( x \) is assumed to be the true value. In practice, \( x \) is usually estimated using a logistic regression model \( x = m(s) \) based on some historical data set. We have conducted some empirical studies and found that the SMR based on estimated \( x \) is substantially right skewed even for moderately large data set. This right-skewness is found to affect the power of test procedures described in the next two sections. We consider
various transformations and found from our empirical study that log(SMR) is much more normal than SMR for a given sample size. A typical example is shown in Figure 2.3. The histogram of SMR still exhibits a substantial right skewness even when the sample size is 5000. Normality is rejected for a small $p$-value of 0.001 for the Shapiro-Wilk test. The histogram of log(SMR) is quite normal with a large $p$-value of 0.3734 for the Shapiro-Wilk test. We thus develop our test procedures using log(SMR) instead of SMR. A similar central limit theorem can be derived for log(SMR) as Theorem 2.

Figure 2.3. Histograms of SMR and log(SMR) with superimposed normal probability density functions. The $p$-values of the Shapiro-Wilk test of normality is 0.001 for SMR and 0.3734 for log(SMR).
Theorem 2. Based on the same assumptions given in Theorem 1, we have

\[ \sqrt{n}(\log(\text{SMR}) - \log(\mu_Y/\mu_X)) \xrightarrow{\text{L}} N\left(0, \frac{\sigma_X^2}{\mu_X^2} - 2\frac{\sigma_{XY}}{\mu_X \mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}\right). \]

Proof. The proof is similar to the proof of Theorem 1.

The asymptotic distribution derived in Theorems 1 and 2 can be used to construct large-sample confidence interval for the true SMR \((\mu_Y/\mu_X)\) in place of Byar’s confidence interval.

§3. Test Procedure I for Comparing SMRs

3.1 Description and Logic of Procedure I

Suppose we want to compare the performances of two surgeons or groups of surgeons based on their SMRs. Let the two surgeons be denoted by \(A\) and \(B\). Assume the true risks of death of patients operated on by surgeons \(A\) and \(B\) are given as \(D_A(s)\) and \(D_B(s)\) respectively. We want to develop a procedure for testing a difference in performance between the two surgeons taking the risk distributions of patients into account. Table 2.1 shows that it may not be meaningful to compare the two SMRs directly with the risks of death of patients estimated using a model based on some national or world-wide data set. This seemingly reasonable approach amounts to comparing \(\mu_{Y_A}/\mu_{X_A}\) and \(\mu_{Y_B}/\mu_{X_B}\) where

\[ \frac{\mu_{Y_A}}{\mu_{X_A}} = \frac{\sum D_A(s)p_A(s)}{\sum x(s)p_A(s)} , \]

and

\[ \frac{\mu_{Y_B}}{\mu_{X_B}} = \frac{\sum D_B(s)p_B(s)}{\sum x(s)p_B(s)} , \]

and \(p_A(s)\) and \(p_B(s)\) are the pmfs of risks of patients operated on by surgeons \(A\) and \(B\) respectively, with risks of death estimated using a model \(x(s) = m(s)\) based on some national or world-wide data set. In general, \(D_A(s)\) or \(D_B(s)\) may not be
the same as \( x(s) \) unless if surgeon \( A \) or \( B \) is a surgeon with performance the same as the average performance of surgeons in the national or world-wide data set. We are interested to compare \( D_A(\cdot) \) and \( D_B(\cdot) \) but if the risk distributions \( p_A(s) \) and \( p_B(s) \) are different, this comparison might lead to an incorrect conclusion as pointed out in the previous section.

We note that if a logistic model is fitted with a surgeon’s data and the model is used to calculate the risk, then the resulting SMR for the surgeon will be 1 irrespective of the risk distribution. For our first procedure, we propose fitting a logistic regression model based on one of the two surgeons’ data, say \( A \). The risks of death of patients operated on by surgeon \( B \) are then estimated using this model for the calculation of surgeon \( B \)’s SMR. If the model is fitted with surgeon \( A \)’s data, surgeon \( A \)’s SMR will be 1 because \( D_A(s) = m(s) = x(s) \). We then have

\[
1 = \frac{\mu_{Y_A}}{\mu_{X_A}} = \frac{\sum D_A(s)p_A(s)}{\sum x(s)p_A(s)} = \frac{\sum D_A(s)p_B(s)}{\sum x(s)p_B(s)},
\]

and hence

\[
E(\text{SMR}_B) \approx \frac{\mu_{Y_B}}{\mu_{X_B}} = \frac{\sum D_B(s)p_B(s)}{\sum x(s)p_B(s)} = \frac{\sum D_B(s)p_B(s)}{\sum D_A(s)p_B(s)}.
\]

Our procedure uses the quantity \( \sum D_B(s)p_B(s) / \sum D_A(s)p_B(s) \) as a measure of difference between surgeons \( A \) and \( B \). It is now meaningful to use \( \text{SMR}_B \) to compare \( D_A(\cdot) \) and \( D_B(\cdot) \) because the risk distribution \( p_B(s) \) is the same. The quantity \( \sum D_B(s)p_B(s) \) is the expected number of deaths for a population with pmf \( p_B(s) \) assuming that they are operated on by surgeon \( B \), while \( \sum D_A(s)p_B(s) \) is the expected number of deaths for the same population assuming that they are operated on by surgeon \( A \). If the quantity \( \sum D_B(s)p_B(s) / \sum D_A(s)p_B(s) \) is larger than 1, surgeon \( A \) is better than surgeon \( B \) since the expected number of deaths operated on by surgeon \( B \) is larger. Similarly, if the quantity \( \sum D_B(s)p_B(s) / \sum D_A(s)p_B(s) \) is smaller than 1, surgeon \( A \) is worse than surgeon \( B \). Therefore, \( \text{SMR}_B \) compares the observed number of deaths for patients operated on by surgeon \( B \) with the
estimates number of deaths if the patients were to be operated on by surgeon A. Therefore, it is meaningful to compare the two surgeons by comparing SMR\textsubscript{B} with 1. Any significant difference has to be due to a difference between \(D_A(\cdot)\) and \(D_B(\cdot)\) since the risk distribution \(p_B(s)\) is the same.

In order to calculate SMR\textsubscript{B}, assume surgeon A operates on \(m\) patients with Parsonnet scores \(S_{A1}, S_{A2}, \cdots, S_{Am}\) taken from a population with pmf \(p_A(s)\) and denote the corresponding surgical outcomes as \(Y_{A1}, Y_{A2}, \cdots, Y_{Am}\). In addition, assume surgeon B operates on \(n\) patients with Parsonnet scores \(S_{B1}, S_{B2}, \cdots, S_{Bn}\) taken from a population with pmf \(p_B(s)\) and denote the corresponding outcomes as \(Y_{B1}, Y_{B2}, \cdots, Y_{Bn}\). We use the data from surgeon A to fit a logistic regression model

\[
\text{logit}(\hat{D}_A(s)) = \hat{\alpha}_A + \hat{\beta}_As, \tag{5}
\]

to estimate the true risk of death of a patient with Parsonnet score \(s\) operated on by surgeon A. Note that the \(\hat{D}_A(s)\) is an estimator of \(D_A(s)\). By using this model, we can estimate the probability of death of a patient operated on by surgeon B assuming the performance of surgeon A. The SMR of surgeon B can then be calculated as

\[
\text{SMR}_B = \frac{\sum_{i=1}^{n} Y_{Bi}}{\sum_{i=1}^{n} \hat{D}_A(S_{Bi})},
\]

where \(\sum_{i=1}^{n} Y_{Bi}\) is the observed number of deaths of patients operated on by surgeon B and \(\hat{D}_A(S_{Bi})\) is the estimated risk of death of a patient with Parsonnet score \(S_{Bi}\) if the patient were to be operated on by surgeon A.

Condition on \(\{(S_{Ai}, Y_{Ai}), i = 1, 2, \cdots, m\}\), we have

\[
E\left(\frac{1}{n} \sum_{i=1}^{n} \hat{D}_A(S_{Bi}) \mid \{ (S_{Ai}, Y_{Ai}), i = 1, 2, \cdots, m\} \right) = \sum \hat{D}_A(s)p_B(s).
\]

In addition,

\[
E\left(\frac{1}{n} \sum_{i=1}^{n} Y_{Bi}\right) = \sum D_B(s)p_B(s),
\]
Using Delta method, it can be shown that
\[
E(\text{SMR}_B) \approx \sum D_B(s)p_B(s)/\sum \hat{D}_A(s)p_B(s).
\]

Note that \(\hat{\alpha}_A\) and \(\hat{\beta}_A\) are MLEs, they are consistent and hence
\[
\hat{D}_A(s) \xrightarrow{p} D_A(s),
\]
\[
\sum \hat{D}_A(s)p_B(s) \xrightarrow{p} \sum D_A(s)p_B(s).
\]

Therefore, as \(m\) increases,
\[
E\left(\frac{1}{n} \sum_{i=1}^{n} \hat{D}_A(S_{Bi})\right) \to \sum D_A(s)p_B(s).
\]

We have
\[
E\left(\frac{1}{n} \sum_{i=1}^{n} \hat{D}_A(S_{Bi})\right) \approx \sum D_A(s)p_B(s),
\]
for large \(m\). If \(\hat{D}_A(s)\) is reasonably accurate for \(D_A(s)\), then \(\text{SMR}_B\) should be approximately an unbiased estimator for the quantity \(\sum D_B(s)p_B(s)/\sum D_A(s)p_B(s)\).

We summarise the calculation of the SMR of surgeon \(B\) using the risks estimated based on a logistic model fitted with data from patients operated on by surgeon \(A\) as the following 3-step procedure:

**Procedure 1: 3-step procedure for calculating SMR\(_B\)**

Step 1. Use the data \(s_{A1}, s_{A2}, \ldots, s_{Am}\) and \(y_{A1}, y_{A2}, \ldots, y_{Am}\) from surgeon \(A\) to fit a logistic regression model
\[
\text{logit}(x_A) = \alpha_A + \beta_A s.
\]

Step 2. Use the model in Step 1 and \(s_{B1}, s_{B2}, \ldots, s_{Bn}\) to estimate the risks of death of patients to be operated on by surgeon \(B\) as \(x_{A,B1}, x_{A,B2}, \ldots, x_{A,Bn}\) assuming the performance of surgeon \(A\).

Step 3. Calculate the SMR for surgeon \(B\) as \(\text{SMR}_B = \frac{\sum_{i=1}^{n} y_{Bi}}{\sum_{i=1}^{n} x_{A,Bi}}\).
The distribution of SMR\textsubscript{B} is difficult to track mathematically, we use the bootstrap method to estimate the distribution. In addition, it is found that the distribution of SMR\textsubscript{B} is still right-skewed even for moderately large sample sizes. Figure 2.3, for example, shows the distribution of SMR\textsubscript{B} based on 5000 simulated samples with \( Q_A = Q_B = 1 \) and two different risk distributions. The distribution of \( \log(\text{SMR}) \) as shown in Figure 2.3 reveals a more normal distribution. The bootstrap method of estimating the quantiles of \( \log(\text{SMR}_B) \) and constructing confidence interval is described below as a four-step procedure:

**Procedure 2: 4-step procedure for estimating the quantiles of \( \log(\text{SMR}_B) \) and constructing confidence interval based on \( \log(\text{SMR}_B) \)**

Step 1. Obtain a bootstrap sample \((s^{*}_{A1}, y^{*}_{A1}), (s^{*}_{A2}, y^{*}_{A2}), \ldots, (s^{*}_{Am}, y^{*}_{Am})\) from \((s_{A1}, y_{A1}), (s_{A2}, y_{A2}), \ldots, (s_{Am}, y_{Am})\) and a bootstrap sample \((s^{*}_{B1}, y^{*}_{B1}), (s^{*}_{B2}, y^{*}_{B2}), \ldots, (s^{*}_{Bn}, y^{*}_{Bn})\) from \((s_{B1}, y_{B1}), (s_{B2}, y_{B2}), \ldots, (s_{Bn}, y_{Bn})\).

Step 2. Use Procedure 1 to calculate \( \log(\text{SMR}_B^*) \) for Surgeon \( B \) using the bootstrap samples obtained in Step 1.

Step 3. Repeat Steps 1 and 2 to get \( N \) bootstrap estimates \( \log(\text{SMR}_B^*) \). Estimate the \( \alpha/2 \) and \( 1 - \alpha/2 \) quantiles of \( \log(\text{SMR}_B) \) as \( \delta \) and \( \bar{\delta} \) respectively based on the \( N \) bootstrap samples.

Step 4. Construct an approximate \((1-\alpha)100\%\) confidence interval for \( \log(\mu_{Y_B}/\mu_{X_B}) \) as \((2\log(\text{SMR}_B) - \bar{\delta}, 2\log(\text{SMR}_B) - \delta)\).

### 3.2 Type I Error and Power of Procedure I

In order to study the probability of type I error and power of Procedure I, we consider simulating data from the real data set \( P_0 \) consisting of 6449 Parsonnet scores and associated outcomes with each operation done by a surgeon from a group of surgeons. Suppose we want to compare two surgeons \( A \) and \( B \) whose performances can be characterized by the odds ratios of death \( Q_A \) and \( Q_B \) respectively. In addition, we assume that surgeon \( A \) operates on patients from a population \( P_A \)
which has the same pmf as $P_0$, and surgeon $B$ operates on patients from a population $P_B$ which has more lower risk patients. The populations $P_A$ and $P_B$ are shown in Table 3.1.

Table 3.1. The populations $P_A$ and $P_B$

<table>
<thead>
<tr>
<th>Population</th>
<th>Low</th>
<th>Elevated</th>
<th>Significantly Elevated</th>
<th>High</th>
<th>Very High</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_A$ or $P_0$</td>
<td>37%</td>
<td>24%</td>
<td>15%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>$P_B$</td>
<td>42%</td>
<td>29%</td>
<td>15%</td>
<td>6%</td>
<td>8%</td>
</tr>
</tbody>
</table>

In our simulation study, we sampled with replacement a Parsonnet score $s$ from population $P_A$ to represent a patient to be operated on by surgeon $A$. The true risk of death of a patient $x_0$ is assumed to follow equation (4). The true risk of death of a patient operated on by surgeon $A$, $x$ is then given as $x/(1-x) = Q_A x_0/(1-x_0)$. To simulate the surgical outcome $y$, a standard uniform random variate was generated and compared with $x$. The simulation of Parsonnet scores, risks of death and surgical outcomes of patients operated on by surgeon $A$ are summarized in the following 4-step procedure:

Procedure 3: 4-step procedure for the simulation of Parsonnet scores, risks of death and surgical outcomes of patients operated on by surgeon $A$

Step 1. Sample randomly 1000 Parsonnet scores $s_{A1}, s_{A2}, \cdots, s_{A1000}$ with replacement from population $P_A$.

Step 2. Calculate the true risks of death assuming the average performance of a surgeon $x_{0,A1}, x_{0,A2}, \cdots, x_{0,A1000}$ using $s_{A1}, s_{A2}, \cdots, s_{A1000}$ and equation (4).

Step 3. Calculate the true risk of death assuming the true performance of surgeon $A$ characterised by the odds of death $Q_A$ as $x_{A,A1}, x_{A,A2}, \cdots, x_{A,A1000}$ using $x_A/(1-x_A) = Q_A x_0/(1-x_0)$.  

153
Step 4. Simulate the surgical outcomes \( y_{A1}, y_{A2}, \cdots, y_{A1000} \) by comparing standard uniform random variates with \( x_{A,A1}, x_{A,A2}, \cdots, x_{A,A1000} \).

Using population \( P_B \) and odds ratio of death \( Q_B \) for surgeon \( B \), Procedure 3 can be employed to simulate similar data for surgeon \( B \): (i) \( s_{B1}, s_{B2}, \cdots, s_{B1000} \), (ii) \( x_{0,B1}, x_{0,B2}, \cdots, x_{0,B1000} \), (iii) \( x_{B,B1}, x_{B,B2}, \cdots, x_{B,B1000} \), (iv) \( y_{B1}, y_{B2}, \cdots, y_{B1000} \). Once the patients’ data are simulated for surgeons \( A \) and \( B \), the SMR of surgeon \( B \) can be calculated using Procedure 1 and the confidence interval based on \( \log(\text{SMR}_B) \) can be calculated using Procedure 2.

We performed the simulation study using \( Q_A = 0.5, 0.8, 1, 1.25, 2 \) and level of significance \( \alpha = 0.05 \). For each value of \( Q_A \), we estimated the true type I error when \( Q_A = Q_B \) and the power for \( Q_B/Q_A \) from 0.5 to 2. Table 3.2 contains the simulated type I errors for various values of \( Q_A \). In general, the estimated type I error is close to specified level of significance. For \( Q_A < 1 \), the test is slightly more conservative than specified but for \( Q_A > 1 \), the test is slightly more liberal. Figure 3.1 shows the power of the test for \( Q_A = 0.5, 0.8, 1, 1.25, 2 \) when \( Q_B/Q_A \) ranges from 0.5 to 2. The power curves show that the proposed test procedure is reasonable because the power is increasing as \( Q_B/Q_A \) deviates more and more from 1.0. For a fixed value of \( Q_B/Q_A \), it can be seen that the power increases as \( Q_A \) increases. This means it is easier to detect a difference between two surgeons if \( Q_A \) is larger.

We can either use surgeon \( A \)’s data to fit a logistic regression model and calculate the SMR for surgeon \( B \) or use surgeon \( B \)’s data to fit a model and calculate the SMR for surgeon \( A \). In order to find out whether there is any difference between the two approaches, we conducted an additional simulation study: we set \( Q_A = 0.5, 0.8, 1, 1.25, 2, Q_B = 0.5, 0.8, 1, 1.25, 2 \), and level of significance \( \alpha = 0.05 \), and calculate the power of the test procedure for all possible pairs of \((Q_A, Q_B)\) for \( Q_A \neq Q_B \), with model fitted using surgeon \( A \)’s data. Note that fitting a model using surgeon \( A \)’s data and calculating \( \text{SMR}_B \) when \( Q_A = 0.8 \) and \( Q_B = 0.5 \) is
equivalent to fitting a model with surgeon B’s data and calculating SMR\textsubscript{A} when \( Q_A = 0.5 \) and \( Q_B = 0.8 \). The power of the test procedure based on SMR\textsubscript{B} is shown in Table 3.3. The percentages of times the two approaches yielded consistent results are shown in Table 3.4. Table 3.3 shows that the powers for testing \( Q_A = Q_B \) are similar whether the logistic model is fitted using surgeon A or surgeon B’s data. Table 3.4 further shows that there is little difference between the two approaches because both produced highly consistent results.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{power_curves.png}
\caption{Plots of power curves for test procedure I corresponding to odds ratio of death \( Q_A = 0.5, 0.8, 1, 1.25, 2 \).}
\end{figure}
Table 3.2. Estimated type I errors (with standard deviations in brackets) for test procedure I corresponding to odds ratio of death $Q_A = Q_B = 0.5, 0.8, 1, 1.25, 2$

<table>
<thead>
<tr>
<th>$Q_A$</th>
<th>0.5</th>
<th>0.8</th>
<th>1</th>
<th>1.25</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I error</td>
<td>0.029</td>
<td>0.045</td>
<td>0.052</td>
<td>0.052</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
</tbody>
</table>

Table 3.3. The powers (with standard deviations in brackets) of test procedure I based on SMR$_B$ for all possible pairs of $(Q_A, Q_B)$ for $Q_A \neq Q_B$ with logistic model fitted using surgeon $A$’s data

<table>
<thead>
<tr>
<th>$Q_A$</th>
<th>$Q_B$</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
<th>1.25</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>–</td>
<td>0.443</td>
<td>0.829</td>
<td>0.985</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>(0.016)</td>
<td>(0.012)</td>
<td>(0.004)</td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>0.8</td>
<td>0.449</td>
<td>–</td>
<td>0.167</td>
<td>0.577</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.016)</td>
<td>–</td>
<td>(0.012)</td>
<td>(0.016)</td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>0.829</td>
<td>0.164</td>
<td>–</td>
<td>0.197</td>
<td>0.977</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>–</td>
<td>(0.013)</td>
<td>(0.005)</td>
<td></td>
</tr>
<tr>
<td>1.25</td>
<td>0.984</td>
<td>0.575</td>
<td>0.183</td>
<td>–</td>
<td>0.794</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td>(0.016)</td>
<td>(0.012)</td>
<td>–</td>
<td>(0.013)</td>
<td></td>
</tr>
<tr>
<td>2.0</td>
<td>1.000</td>
<td>0.998</td>
<td>0.972</td>
<td>0.792</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.001)</td>
<td>(0.005)</td>
<td>(0.013)</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>
Table 3.4. The percentages of times (with standard deviations in brackets) the two approaches of test procedure I yielding the same result

<table>
<thead>
<tr>
<th>$Q_B$</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
<th>1.25</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>98.1</td>
<td>93.2</td>
<td>95.0</td>
<td>98.9</td>
<td>100.0</td>
</tr>
<tr>
<td></td>
<td>(0.4)</td>
<td>(0.8)</td>
<td>(0.7)</td>
<td>(0.3)</td>
<td>(0.000)</td>
</tr>
<tr>
<td>0.8</td>
<td>–</td>
<td>98.5</td>
<td>94.5</td>
<td>91.8</td>
<td>99.9</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>(0.4)</td>
<td>(0.7)</td>
<td>(0.9)</td>
<td>(0.1)</td>
</tr>
<tr>
<td>$Q_A$</td>
<td>1.0</td>
<td>–</td>
<td>–</td>
<td>97.9</td>
<td>95.4</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(0.5)</td>
<td>(0.7)</td>
</tr>
<tr>
<td>1.25</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>98.7</td>
<td>93.8</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(0.4)</td>
<td>(0.8)</td>
</tr>
<tr>
<td>2.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>98.1</td>
</tr>
<tr>
<td></td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(0.4)</td>
</tr>
</tbody>
</table>

3.3 Effect of Sample Size

In order to investigate the effect of the sample size on the test procedure, we conducted a further simulation study to assess the type I error of the procedure for surgeons with various odds ratio of death with respect to different sample sizes. We set the odds ratio of mortality of the surgeons to be $Q_A = Q_B = 0.5, 0.8, 1, 1.25, 2$ and each surgeon operating on 500, 600, 700, 800, 900 and 1000 patients. The level of significance of the procedure is set at $\alpha = 0.05$ for each case. Table 3.5 shows the observed type I errors for the various cases.

In general, the observed type I error is closer to the specified level of significance for larger sample sizes. In addition, as $Q_A$ increases, the observed type I error becomes closer to the specified level of significance. The discrepancy is the largest when sample size and $Q_A$ are both small. This is due to an increase in the variance because of poorly fitted logistic models based on small number of deaths. This is not unexpected because the quality of any procedure that is based on a logistic model depends greatly on fitting a satisfactory logistic model.
Table 3.5. Observed type I errors (with standard deviations in brackets) of test procedure I for surgeons with various odds ratio of death with respect to different sample sizes and level of significance set at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$Q_A = Q_B$</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
<th>1.25</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(0.003)</td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>500</td>
<td></td>
<td>0.010</td>
<td>0.031</td>
<td>0.029</td>
<td>0.023</td>
<td>0.039</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>600</td>
<td></td>
<td>0.014</td>
<td>0.031</td>
<td>0.034</td>
<td>0.038</td>
<td>0.047</td>
</tr>
<tr>
<td>700</td>
<td></td>
<td>0.014</td>
<td>(0.029)</td>
<td>0.036</td>
<td>0.037</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>800</td>
<td></td>
<td>0.032</td>
<td>0.035</td>
<td>0.037</td>
<td>0.043</td>
<td>(0.044)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>900</td>
<td></td>
<td>0.038</td>
<td>0.039</td>
<td>0.040</td>
<td>0.056</td>
<td>0.053</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.029</td>
<td>0.045</td>
<td>0.052</td>
<td>0.052</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
</tbody>
</table>

§4. Test Procedure II for Comparing SMRs

4.1 Description and Logic of Procedure II

In this section, we develop a second procedure which is based on fitting a logistic model using the combined data of two surgeons. The resulting model estimates the risk based on the average performance of the two surgeons. Assume the true risk of death of a patient operated by a surgeon with the average performance of the two surgeons can be characterized by

$$\text{logit}(x_c) = \alpha_c + \beta_c s.$$  \hspace{1cm} (6)

If the two surgeons have the same performance, then $D_A(s) = D_B(s) = x_c(s)$ and hence

$$E(SMR_A) \approx \mu_{Y_A}/\mu_{X_cA} = \sum D_A(s)p_A(s)/\sum x_c(s)p_A(s) = 1,$$
and
\[ E(SMR_B) \approx \frac{\mu_{Y_B}}{\mu_{X_{cB}}} = \sum D_B(s)p_B(s)/\sum x_c(s)p_B(s) = 1, \]
where \( \mu_{X_{cA}} \) and \( \mu_{X_{cB}} \) are the true mean risks of death assuming the average performance of the two surgeons for populations \( P_A \) and \( P_B \) respectively. Hence, \( \log(\mu_{Y_B}/\mu_{X_{cB}}) - \log(\mu_{Y_A}/\mu_{X_{cA}}) = 0 \). If the two surgeons do not have the same performance, model (6) will overestimate a patient’s risk operated by the better surgeon and underestimate for the other surgeon. This means that one of the two quantities \( \mu_{Y_A}/\mu_{X_{cA}} \) and \( \mu_{Y_B}/\mu_{X_{cB}} \) associated with the better surgeon will be smaller than 1 and the other will be greater than 1. Then, \( \log(\mu_{Y_B}/\mu_{X_{cB}}) - \log(\mu_{Y_A}/\mu_{X_{cA}}) \neq 0 \). Therefore, it is reasonable to compare the two surgeons using the statistic \( \log(SMR_A) - \log(SMR_B) \). In fact, the statistic \( SMR_A - SMR_B \) is also possible but the power of this statistic is found to be greatly affected by the right skewness of SMR. We thus consider only \( \log(SMR_A) - \log(SMR_B) \). The following 3-step procedure can be used to calculate the SMRs of two surgeons using the risks estimated using a model fitted with the combined patients’ data.

**Procedure 4: 3-step procedure for calculating SMR\(_A\) and SMR\(_B\)**

Step 1. Use the combined data \((s_{A1}, y_{A1}), (s_{A2}, y_{A2}), \ldots, (s_{Am}, y_{Am})\) and \((s_{B1}, y_{B1}), (s_{B2}, y_{B2}), \ldots, (s_{Bn}, y_{Bn})\) to fit a logistic model (6).

Step 2. Use model (6) and \( s_{A1}, s_{A2}, \ldots, s_{Am} \) to estimate the risks of death for patients operated on by surgeon \( A \) as \( x_{c,A1}, x_{c,A2}, \ldots, x_{c,Am} \) and calculate
\[ SMR_A = \sum y_{Ai}/\sum x_{c,Ai}. \]

Step 3. Use model (6) and \( s_{B1}, s_{B2}, \ldots, s_{Bn} \) to estimate the risks of death for patients operated on by surgeon \( B \) as \( x_{c,B1}, x_{c,B2}, \ldots, x_{c,Bn} \), and calculate
\[ SMR_B = \sum y_{Bi}/\sum x_{c,Bi}. \]

Let \( U = \log(SMR_B) - \log(SMR_A) \). Similar to the test statistic of test procedure I, the distribution of \( U \) is also difficult to track mathematically, we thus use the
bootstrap method to estimate the distribution of $U$. The bootstrap method of estimating the quantiles of $U$ and constructing confidence interval is given as a 4-step procedure:

**Procedure 5: 4-step procedure for estimating the quantiles of $U$ and constructing confidence interval based on $U$**

Step 1. Obtain a bootstrap sample $(s_{A1}^*, y_{A1}^*)$, $(s_{A2}^*, y_{A2}^*), \ldots, (s_{Am}^*, y_{Am}^*)$ from $(s_A, y_A)$, $(s_{A2}, y_{A2}), \ldots, (s_{Am}, y_{Am})$ and a bootstrap sample $(s_{B1}^*, y_{B1}^*)$, $(s_{B2}^*, y_{B2}^*), \ldots, (s_{Bn}^*, y_{Bn}^*)$ from $(s_B, y_B)$, $(s_{B2}, y_{B2}), \ldots, (s_{Bn}, y_{Bn})$.

Step 2. Use Procedure 4 to calculate the bootstrap estimate $U^*$ using the bootstrap samples obtained in Step 1.

Step 3. Repeat Steps 1 and 2 to get $N$ bootstrap estimates $U^*$ and estimate the $\alpha/2$ and $1 - \alpha/2$ quantiles of $U$ as $\hat{\delta}$ and $\tilde{\delta}$ respectively based on the $N$ bootstrap samples.

Step 4. Construct an approximate $(1-\alpha)100\%$ confidence interval for $\log(\mu_{Y_B}/\mu_{XeB}) - \log(\mu_{Y_A}/\mu_{XeA})$ as $(2U - \tilde{\delta}, 2U - \hat{\delta})$.

### 4.2 Type I Error and Power of Procedure II

We conducted a simulation study to assess the type I error and power of procedure II. We use the same assumptions given in the simulation study for procedure I, namely surgeons $A$ and $B$ operates on patients from population $P_A$ and $P_B$ respectively, and the performance of surgeons $A$ and $B$ are characterized by the odds ratio of death $Q_A$ and $Q_B$ respectively. The true risk of death of a patient $x_0$ is assumed to follow equation (4). The true risk of death of a patient operated on by a surgeon is determined using $x/(1 - x) = Qx_0/(1 - x_0)$, where $Q = Q_A$ for surgeon $A$ and $Q = Q_B$ for surgeon $B$. To simulate the surgical outcome $y$, a standard uniform random variate was generated and compared with $x$. The simulation of Parsonnet scores, risks of death and surgical outcomes of patients operated on by surgeons $A$ and $B$ can be done using Procedure 3. Once the patients’ data are simulated
for surgeons $A$ and $B$, SMR$_A$ and SMR$_B$ can be calculated using Procedure 4 and the confidence interval based on $\log(\text{SMR}_B) - \log(\text{SMR}_A)$ can be calculated using Procedure 5.

We performed the simulation study for $Q_A = 0.5, 0.8, 1, 1.25, 2$ using the level of significance $\alpha = 0.05$. For each value of $Q_A$, we estimated the true type I error for $Q_A = Q_B$ and the power for $Q_B/Q_A$ from 0.5 to 2. Table 4.1 contains the simulated type I errors for the 5 values of $Q_A$. In general, the estimated type I error is close to the underlying level of significance.

*Figure* 4.1 shows the power of test procedure II for $Q_A = 0.5, 0.8, 1, 1.25, 2$ with $Q_B/Q_A$ ranges from 0.5 to 2. The power curves show that procedure II is also reasonable since the power is increasing as $Q_B/Q_A$ deviates more and more from 1.0. For a fixed value of $Q_B/Q_A$, the power increases as $Q_A$ increases. This means it is easier to detect a difference between two surgeons if $Q_A$ is larger.
Table 4.1. Estimated type I errors (with standard deviations in brackets) for test procedure II corresponding to odds ratio of death $Q_A = Q_B = 0.5, 0.8, 1, 1.25, 2$

<table>
<thead>
<tr>
<th>$Q_A$</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
<th>1.25</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I Error</td>
<td>0.041</td>
<td>0.045</td>
<td>0.043</td>
<td>0.049</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
</tbody>
</table>

Figure 4.1. Plots of power curves for test procedure II corresponding to odds ratio of death $Q_A = 0.5, 0.8, 1, 1.25, 2$. 

162
Table 4.2. Observed type I errors (with standard deviations in brackets) of test procedure II for surgeons with various odds ratios of death with respect to different sample sizes and level of significance set at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>Sample size</th>
<th>$Q_A = Q_B$</th>
<th>0.5</th>
<th>0.8</th>
<th>1.0</th>
<th>1.25</th>
<th>2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td></td>
<td>0.020</td>
<td>0.024</td>
<td>0.049</td>
<td>0.036</td>
<td>0.038</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.004)</td>
<td>(0.005)</td>
<td>(0.007)</td>
<td>(0.006)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>600</td>
<td></td>
<td>0.026</td>
<td>0.032</td>
<td>0.038</td>
<td>0.034</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>700</td>
<td></td>
<td>0.024</td>
<td>0.026</td>
<td>0.044</td>
<td>0.048</td>
<td>0.046</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>800</td>
<td></td>
<td>0.022</td>
<td>0.030</td>
<td>0.038</td>
<td>0.056</td>
<td>0.040</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.005)</td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
<tr>
<td>900</td>
<td></td>
<td>0.030</td>
<td>0.034</td>
<td>0.046</td>
<td>0.045</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.005)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.007)</td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td>0.041</td>
<td>0.045</td>
<td>0.043</td>
<td>0.049</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.006)</td>
</tr>
</tbody>
</table>
4.3 Effect of Sample Size

In order to investigate the effect of the sample size on test procedure II, we conducted a simulation study to evaluate type I error of the procedure for surgeons with various odds ratios of mortality with respect to different sample sizes. We set the odds ratios of mortality of the surgeons to be \( Q_A = Q_B = 0.5, 0.8, 1, 1.25, 2 \) and each operating on 500, 600, 700, 800, 900 and 1000 patients. The level of significance of the procedure is set at \( \alpha = 0.05 \) for each case. Table 4.2 contains the observed type I errors for the different situations. In general, the observed type I error is closer to the specified level of significance for larger sample sizes. In addition, as \( Q_A \) increases, the observed type I error becomes closer to the specified level of significance. The discrepancy is the largest when sample size and \( Q_A \) are both small. This is due to an increase in the variance because of poorly fitted logistic models based on small number of deaths.

4.4. Comparison with Test Procedure I

The power curves of test procedures I and II for odds ratio of death \( Q_A = 0.5, 0.8, 1, 1.25 \) and 2 are displayed in Figures 4.2–4.6. It is clear from these figures that the two procedures have very similar power performance.

![Figure 4.2](image)

*Figure 4.2. Plots of power curves for procedure I (dashed line) and II (solid line) for odds ratio \( Q_A = 0.5 \)*
Figure 4.3. Plots of power curves for procedure I (dashed line) and II (solid line) for odds ratio $Q_A = 0.8$

Figure 4.4. Plots of power curves for procedure I (dashed line) and II (solid line) for odds ratio $Q_A = 1$

Figure 4.5. Plots of power curves for procedure I (dashed line) and II (solid line) for odds ratio $Q_A = 1.25$
In this section, we compare the performances of test procedures I and II based on 3 simulated and 2 real data sets. In order for the 3 simulated data sets to be representative of real situations, we consider taking random samples of Parsonnet scores from the real population of Parsonnet scores, $P_0$ as described in Section 2. We let surgeon $A$ operate on patients taken from the population $P_A$ which is the same as $P_0$. The population $P_{B1}$ is also taken to be the same as $P_0$. The other two populations $P_{B2}$ and $P_{B3}$ are created from $P_A$ such that $P_{B2}$ has more low risk patients than $P_A$ while $P_{B3}$ has less low risk patients than $P_A$. The risk distributions of the 4 populations are shown in Table 5.1. Surgeon $B$ operates on patients taken from $P_{B1}$, $P_{B2}$ and $P_{B3}$. The different risk distributions allow their effect on the test procedures to be studied. The true risk of death of a patient $x_0$ is assumed to follow model (4). The odds ratio of death associated with surgeons $A$ and $B$ are fixed at $Q_A = 1$ and $Q_B = 1.25$ respectively. The true risk of death $x$ is then determined as $x/(1-x) = Qx_0/(1-x_0)$ where $Q$ is either $Q_A$ or $Q_B$. For each case, samples of 1000 Parsonnet scores were randomly selected with replacement from the two populations.
Table 5.1. Risk distributions of 4 populations of Parsonnet scores

<table>
<thead>
<tr>
<th>Population</th>
<th>Low</th>
<th>Elevated</th>
<th>Significantly elevated</th>
<th>High</th>
<th>Very high</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>37%</td>
<td>24%</td>
<td>15%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>$P_A$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_{B1}$</td>
<td>37%</td>
<td>24%</td>
<td>15%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>$P_{B2}$</td>
<td>42%</td>
<td>29%</td>
<td>15%</td>
<td>6%</td>
<td>8%</td>
</tr>
<tr>
<td>$P_{B3}$</td>
<td>32%</td>
<td>19%</td>
<td>15%</td>
<td>16%</td>
<td>18%</td>
</tr>
</tbody>
</table>

Table 5.2. Confidence intervals calculated using procedures I, II and Byar’s procedure for 2 real data sets and 3 simulated data sets generated from a real population of Parsonnet scores

<table>
<thead>
<tr>
<th>Case</th>
<th>$A$</th>
<th>$B$</th>
<th>Procedure I</th>
<th>Procedure II</th>
<th>Byar’s procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$P_A$</td>
<td>$P_{B1}$</td>
<td>(0.09, 0.71)</td>
<td>(0.09, 0.70)</td>
<td>(0.67, 1.13) (1.06, 1.63)</td>
</tr>
<tr>
<td>2</td>
<td>$P_A$</td>
<td>$P_{B2}$</td>
<td>(0.04, 0.61)</td>
<td>(0.03, 0.68)</td>
<td>(0.60, 1.04) (0.92, 1.50)</td>
</tr>
<tr>
<td>3</td>
<td>$P_A$</td>
<td>$P_{B3}$</td>
<td>(0.04, 0.62)</td>
<td>(0.05, 0.62)</td>
<td>(0.73, 1.20) (1.06, 1.58)</td>
</tr>
<tr>
<td>4</td>
<td>$S_{A1}$</td>
<td>$S_{B1}$</td>
<td>(-1.08, -0.13)</td>
<td>(-0.95, -0.10)</td>
<td>(1.05, 1.44) (0.44, 1.05)</td>
</tr>
<tr>
<td>5</td>
<td>$S_{A2}$</td>
<td>$S_{B2}$</td>
<td>(-1.00, -0.06)</td>
<td>(-0.98, 0.04)</td>
<td>(0.80, 2.14) (0.67, 1.22)</td>
</tr>
</tbody>
</table>

Table 5.2 contains the 95% confidence intervals calculated using procedures I and II. For procedure I, the confidence interval is based on $\log(\text{SMR}_B)$ with the risks estimated using a logistic model fitted with surgeon $A$’s data. For procedure II, the confidence interval is based on $\log(\text{SMR}_B) - \log(\text{SMR}_A)$ with the risks estimated using a logistic model fitted with the combined data. Both procedures indicate a significant difference between surgeons $A$ and $B$. The Byar’s confidence intervals are also constructed for the two surgeons. As pointed out in Section 1, this confidence interval may not be valid unless the estimated number of deaths is close to the expected value. It is however interesting to note that for each case, the two Byar’s confidence intervals overlap and hence dose not provide any evidence of a significant
difference. This is probably due to the wider limits as shown by Faris et al. (2003) in a simulation study.

For the first real data set, the histograms of samples of 1654 Parsonnet scores \(S_{A1}\) and 781 Parsonnet scores \(S_{B1}\) are displayed in Figure 5.1. A comparison of the two histograms shows that relatively more low-risk patients were operated on by surgeon \(B\). The confidence intervals calculated using procedures I and II and Byar’s procedure are displayed in Table 5.2. Both procedures I and II show that Surgeon \(B\) is significantly better than Surgeon \(A\). On the contrary, Byar’s procedure does not indicate any significant difference.

For the second real data set, the histograms of samples of 202 Parsonnet scores \(S_{A2}\) and 609 Parsonnet scores \(S_{B2}\) are displayed in Figure 5.2. The confidence intervals calculated using procedures I and II and Byar’s procedure are displayed in Table 5.2. Procedure I shows that Surgeon \(B\) is significantly better than Surgeon \(A\) even though for this case, surgeon \(A\) operated on more low-risk patients. Procedure II is almost significant at the 0.05 level while Byar’s procedure does not indicate any significant difference.
Figure 5.1. The histograms of Parsonnet scores of samples $S_{A1}$ and $S_{B1}$.
§6. Conclusion

A new application of the traditional SMR in measuring the performance of a surgeon or a group of surgeons results in a SMR which compares an observed number of deaths with an estimated number of deaths based on the average performance of surgeons contained in some national or world-wide data set. The estimated number depends not only on the sample size but also on the risk distributions of patients. It is meaningful to use the SMR to compare the performance of a surgeon with a surgeon of average performance according to the historical data set. However, it may not be meaningful to compare the performances of two surgeons based on their SMRs when both SMRs are above 1 or below 1. We show that a better surgeon can have a higher SMR if the surgeon operates on patients of lower
risks. We also show that the SMRs of two equally competent surgeons are not necessarily the same because the SMR is influenced partly by the risk distribution of the patients. Most of the published SMRs are calculated using estimated number of deaths based on some national or world-wide data set. Inevitably, the SMRs are naively compared to provide a so-called “informed” choice. In order to compare two surgeons scientifically, we developed two procedures based on fitting a logistic model using one of the two surgeons’ data or the combined data. The two procedures are similar in terms of type I error and power. All existing confidence interval derived based on an expected number of deaths may not be valid for the new SMR because they assume the estimated number of deaths is a constant. For this reason, the Byar’s confidence interval was not compared with our procedures in a comprehensive manner. Nonetheless, there is some empirical evidence to show that our procedures are more powerful even if the estimated number of deaths can be assumed to be close to the expected number of deaths.
Appendix A: R Codes for Procedure I

1 ########################################This is procedure I###################################

2

3 ################################################################################Simulation ################################################################################

4 sgh=read.table("SGH1_data.txt");
5 record=sgh[,13];parsonnet=sgh[,15];
6 nrecord=length(record);
7 record1=record2=rep(0,nrecord);
8 for(i in 1:nrecord)
9 {if(record[i]<31&record[i]>-1) record1[i]=1 else record2[i]=1}
10 record3=cbind(record1,record2);
11 logitm(glm(record3~parsonnet,family=binomial("logit"));
12 po1=parsonnet;po2=read.table("po2.txt");po2=po2[,1];
13 population1=po1;population2=po2;
14
15 ################################################################################get the true risk of patients################################################################################
16 getrisk=function(parsonnetscore,Odds){
17 logitp=predict(logitm,data.frame(parsonnet=parsonnetscore));
18 risk=exp(logitp)/(1+exp(logitp));
19 truerisk=Odds*risk/(1-risk+Odds*risk);
20 return(truerisk);
21 }
22
23 ################################################################################get outcome for patients################################################################################
24 getresult=function(parsonnetscore,Odds){
25 truerisk=getrisk(parsonnetscore,Odds);
26 number=length(parsonnetscore);
27 test=runif(number);
28 result=rep(0,number);
29 for(i in 1:number) {if(test[i]<truerisk[i]) result[i]=1}
30 return(result);
31 }
32
33 ################################################################################calculate log SMR for two sample################################################################################
34 getlogsmr=function(parsam1,outcome1,parsam2,outcome2){
35 if(var(outcome1)==0) logSMR=1000000
36 return(logSMR)
37 }
else{
    record1=cbind(outcome1,1-outcome1);
    modelt=glm(record1~parsam1,family=binomial("logit"));
    logitt=predict(modelt,data.frame(parsam1=parsam2));
    riskt=exp(logitt)/(1+exp(logitt));
    logSMR=log(mean(outcome2)/mean(riskt));
}
return(logSMR);

# get variance for log SMR by bootstrap
getquantile=function(parsam1,outcome1,parsam2,outcome2,numberofbootstrap){
    tn=numberofbootstrap;
    le=length(parsam1);
    index=(1:le);
    record=rep(0,tn);
    for(i in 1:tn){
        indexsam1=sample(index,le,TRUE);
        bootparsam1=parsam1[indexsam1];
        bootoutcome1=outcome1[indexsam1];
        indexsam2=sample(index,le,TRUE);
        bootparsam2=parsam2[indexsam2];
        bootoutcome2=outcome2[indexsam2];
        record[i]=getlogsmr(bootparsam1,bootoutcome1,bootparsam2,bootoutcome2);
    }
    var=var(record);
    q=c(0.025,0.975);
    quantiles=quantile(record,q);
    return(quantiles);
}

# check for cover
getcover=function(samplesize1,Odds1,samplesize2,Odds2,numberofbootstrap){
  sample1=sample(population1,samplesize1,TRUE);
  sample2=sample(population2,samplesize2,TRUE);
  outcome1=getresult(sample1,Odds1);
  outcome2=getresult(sample2,Odds2);
  logSMR=getlogsmr(sample1,outcome1,sample2,outcome2);
  quantiles=getquantile(sample1,outcome1,sample2,outcome2,numberofbootstrap);
  l=2*logSMR-quantiles[2];
  u=2*logSMR-quantiles[1];
  if(l<0&u>0) mark=1 else mark=0
  return(mark);
}

getcoverage=function(samplesize1,Odds1,samplesize2,Odds2,
                    numberofbootstrap,numberoftrials){
  marks=rep(0,numberoftrials);
  for(i in 1:numberoftrials)
    marks[i]=getcover(samplesize1,Odds1,samplesize2,Odds2,numberofbootstrap);
  coverage=mean(marks);
  return(coverage);
Appendix B: R Codes for Procedure II

###This is procedure II

```
####Simulation

sgh = read.table("C:/Users/g0900761/Documents/SGH1_data.txt");
record = sgh[,13]; parsonnet = sgh[,15];
nrecord = length(record);
record1 = record2 = rep(0, nrecord);
for (i in 1:nrecord)
  {if (record[i] < 31 & record[i] > -1) record1[i] = 1 else record2[i] = 1}
record3 = cbind(record1, record2);
logitm = glm(record3 ~ parsonnet, family = binomial("logit"));
po1 = parsonnet; po2 = read.table("po2.txt"); po2 = po2[,1];
population1 = po1; population2 = po2;

###get the true risk of patients

getrisk = function(parsonnetscore, Odds){
  logitp = predict(logitm, data.frame(parsonnet = parsonnetscore));
  risk = exp(logitp)/(1+exp(logitp));
  truerisk = Odds*risk/(1-risk+Odds*risk);
  return(truerisk);
}

###get outcome for patients

getresult = function(parsonnetscore, Odds){
  truerisk = getrisk(parsonnetscore, Odds);
  number = length(parsonnetscore);
  test = runif(number);
  result = rep(0, number);
  for (i in 1:number) {if (test[i] < truerisk[i]) result[i] = 1}
  return(result);
}

###calculate difference of log SMR for two sample

getdlogsmr = function(parsam1, outcome1, parsam2, outcome2){
  pooledpar = c(parsam1, parsam2);
```
pooledoutcome=c(outcome1,outcome2);

if(var(pooledoutcome)==0) dlogsmr=0
else{
pooledrecord=cbind(pooledoutcome,1-pooledoutcome);
modelt=glm(pooledrecord~pooledpar,family=binomial("logit"));

logitt1=predict(modelt,data.frame(pooledpar=parsam1));
riskt1=exp(logitt1)/(1+exp(logitt1));
logsmr1=log(mean(outcome1)/mean(riskt1));

logitt2=predict(modelt,data.frame(pooledpar=parsam2));
riskt2=exp(logitt2)/(1+exp(logitt2));
logsmr2=log(mean(outcome2)/mean(riskt2));

dlogsmr=logsmr1-logsmr2;
}

return(dlogsmr);

getquantile=function(parsam1,outcome1,parsam2,outcome2,numberofbootstrap){
  tn=numberofbootstrap;
  le1=length(parsam1);
  index1=(1:le1);
  le2=length(parsam2);
  index2=(1:le2);
  record=rep(0,tn);
  for(i in 1:tn){
    indexsam1=sample(index1,le1,TRUE);
    bootparsam1=parsam1[indexsam1];
    bootoutcome1=outcome1[indexsam1];
    indexsam2=sample(index2,le2,TRUE);
    bootparsam2=parsam2[indexsam2];
    }
bootoutcome2=outcome2[indexsam2];
record[i]=getdlogsmr(bootparsam1,bootoutcome1,bootparsam2,bootoutcome2); }

var=var(record);
q=c(0.025,0.975);
quantiles=quantile(record,q);
return(quantiles);

getcover=function(samplesize1,Odds1,samplesize2,Odds2,numberofbootstrap){
sample1=sample(population1,samplesize1,TRUE);
sample2=sample(population2,samplesize2,TRUE);
outcome1=getresult(sample1,Odds1);
outcome2=getresult(sample2,Odds2);
dlogSMR=getdlogsmr(sample1,outcome1,sample2,outcome2);
quantiles=getquantile(sample1,outcome1,sample2,outcome2,numberofbootstrap);
l=2*dlogSMR-quantiles[2];
u=2*dlogSMR-quantiles[1];
if(l<0&u>0) mark=1 else mark=0
return(mark);
}

getcoverage=function(samplesize1,Odds1,samplesize2,Odds2,
numberofbootstrap,numberoftrials){
marks=rep(0,numberoftrials);
for(i in 1:numberoftrials)
marks[i]=getcover(samplesize1,Odds1,samplesize2,Odds2,numberofbootstrap);
coverage=mean(marks);
return(coverage);
References


Canadian Institute for Health Information (2012), *Technical Notes: Hospital Standardized Mortality Ratio (HSMR)*. Ottawa, Canada: The Institute.


