RISK ADJUSTMENT IN CLINICAL PROCEDURES

LOKE CHOK KANG

(B.Sci.(Hons.), NUS)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY

NATIONAL UNIVERSITY OF SINGAPORE

2010

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my heartfelt gratitude to the following people:

to **my supervisor, Associate Professor Gan Fah Fatt,**

for his patience, guidance and suggestions,

without which this dissertation would definitely

not have been possible;

to **Dr Andy Chiang, Professor Loh Wei-Liem, Ms Yvonne Chow,**

Mr Zhang Rong, Ms Zhang Rongli, Ms Lee Huey Chyi

Ms Wong Yean Ling, Associate Professor Chua Tin Chiu,

for their invaluable advice and help given;

to **my parents,**

for their encouragement, meticulous care and

love that they showered upon me;

to **NUS research grant (No. R155-000-092-112) for the project,**

"Risk-Adjusted Cumulative Sum Control Charting Procedures",

for the support and assistance in my PhD program;

a very big **THANK YOU** to all of you and many others.

2010

TABLE OF CONTENTS

SUMMARY

The evolution of the assessment of medical practice has been speeding up tremendously, as seen from recent literature (discussed in later chapters). However, patients in hospitals tend to differ notably in terms of mortality risk. This variability might result in additional fluctuation in the outcomes, thus masking the effectiveness, and resulting in misapprehension of the true assessment. In this dissertation, a systematic approach to assess clinical procedures is taken by taking into account this variability in the mortality risk and subsequently focusing on three major areas: statistical process control, comparison of procedures, and overall quality indicators.

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. GENERAL INTRODUCTION

Section 1. Introduction

The evolution of the assessment of medical practice has been speeding up tremendously, as seen from recent literature (discussed in later chapters). However, realistically in an industrial setting where the raw materials or products may be comparably homogeneous in nature, this is dissimilar to that for the health care delivery. Patients in hospitals tend to differ notably in terms of pre-procedural risk of failure, which in this dissertation, we will refer to as mortality risk. If this variability in the mortality risk is not taken into account in the assessment of medical practice, this variability might result in additional fluctuation in the outcomes, thus masking the effectiveness, and resulting in misapprehension of the true situation. Due to this variability, it does not make sense to discuss the assessment of medical practice without first accounting for risk adjustment. Motivated by the above discussion, the focus of this dissertation is on risk adjustment in clinical procedures.

Section 2. Dissertation Organization

This dissertation is organized using the "alternative format" of compiling together several manuscripts prepared for submission to international journals. For the assessment of clinical procedures, this dissertation takes a systematic approach to assess clinical procedures by focusing on three major areas: statistical process control, comparison of procedures, and overall quality indicators.

Chapter 2 utilizes the fundamental techniques of statistical process control through the introduction of risk-adjusted monitoring tools. At present, risk-adjusted monitoring tools are only used to monitor clinical performances. But we demonstrate that it is not sufficient to solely monitor clinical performances. As such, a joint monitoring scheme for clinical performance and the mortality risk is proposed. This scheme is not just necessary but also essential to avoid making erroneous inferences on clinical performance when the risk distribution has changed. A new charting procedure to monitor the mortality risk distribution, specifically the average mortality risk of patients, is also introduced.

At present, risk-adjusted analytical tools are best used as a monitoring procedure, rather than to compare clinical performances. In Chapter 3, we propose a model-free diagnostic technique to estimate the actual mortality rates for all levels of predicted mortality risk to assess clinical performances. Using these estimated mortality rates, we present a set of risk-adjusted test procedures which alleviate the problem of interpretation through the use of penalty-reward scores. We also consider other risk-adjusted methods for this comparison.

One widely-used overall quality indicator in medical practice will be the standardized mortality ratio (SMR). However, despite being around for some time, health service providers are still skeptical on its ability to truly identify poor-quality providers. Chapter 4 will present various limitations of using the SMR, as well as highlight various possibly wrong interpretations through the use of SMR. Chapter 5 contains a general conclusion for the dissertation.

CHAPTER 2: JOINT MONITORING SCHEME FOR CLINICAL PERFORMANCES AND MORTALITY RISK

SUMMARY

Measuring quality of medical practice is a key component in improving efficiency in health care, such assessment is playing an increasingly prominent role in quality management. At present, risk-adjusted monitoring tools are only used to monitor clinical performances. Using a sensitivity analysis, as well as illustrations using real life applications and simulated examples, we demonstrated that it is not sufficient to solely monitor clinical performances. In this paper, we propose to jointly monitor clinical performance and the mortality risk. This joint monitoring is not just necessary but also essential to avoid making erroneous inferences on clinical performance when the risk distribution has changed. We also proposed a new charting procedure to monitor the mortality risk distribution, specifically the average mortality risk of patients. The design of the joint monitoring scheme is also described in detail, with an illustration based on a real data set.

SECTION 1. INTRODUCTION

The evolution of the assessment of medical practice has been speeding up tremendously, as seen from recent literature (Werner and Bradlow, 2006, Clarke and Oakley, 2007, Krumholz et al., 2008, Biswas and Kalbfleisch, 2008, Steiner and Jones, 2009). Measuring quality of medical practice is a key component in improving efficiency in health care, such assessment is playing an increasingly prominent role in quality management. One fundamental practice of assessment will be that of clinical performance monitoring. In 1999, an independent body, the UK National Institute of Clinical Excellence was established, after the UK General Medical Council found three doctors possibly guilty of professional misconduct over the quality of their heart surgeries conducted. The professional misconduct led to 29 mortalities out of 53 children who were operated at the Bristol Royal Infirmary (2001, BBC News 1998). This depicts the importance of clinical performance monitoring as timely signals of deteriorated performance can be used to identify assignable causes and this will in turn avoid future avertible mortalities or other adverse health issues.

Monitoring of the effectiveness of clinical procedures and physicians' performance has been popularized well over 50 years ago in the medical field (Armitage, 1954 and Bartholomay, 1957). Other works include Chen (1978), Kenett and Pollak (1983), Gallus et al. (1986), Frisen and De Mare (1991), Frisen (1992), Chen (1996), Rossi, Lampugnani and Marchi (1999), Steiner, Cook and Farewell (1999), Steiner et al. (2000), Spiegelhalter et al. (2003), Cook et al. (2003), Grigg and Farewell (2004), Sherlaw-Johnson et al. (2005), Sherlaw-Johnson, Wilson and

Gallivan (2007), Grigg and Spiegelhalter (2007), Biswas and Kalbfleisch (2008), Steiner and Jones (2009), and Gan and Tan (2010). These works show the increasing importance and popularity of such monitoring schemes in the health care industry as it is fundamental that the quality of service provided by health care providers are consistent and acceptable.

Realistically in an industrial setting where the raw materials or products may be comparably homogeneous in nature, this is dissimilar to that for the health care delivery. Patients in hospitals tend to differ notably in terms of pre-procedural risk of failure, which in this paper we will refer to as mortality risk. If this variability in the mortality risk is not taken into account when assessing the effectiveness of a certain clinical procedure, this variability might result in additional fluctuation in the outcomes, thus masking the effectiveness, and resulting in misapprehension of the true situation. Due to this variability, it does not make sense to monitor clinical performance without risk adjustment because the physician or clinical procedure which was only conducted on patients with high risks will tend to have a significant lower success rate. It is therefore sensible to monitor clinical performance while accounting for the mortality risk of patients.

Due to the necessity for risk adjustment, Lovegrove et al. (1997, 1999) and Poloniecki, Valencia and LittleJohns (1998) proposed a simple monitoring scheme, the variable life-adjusted display (VLAD) which plots the expected mortality count subtracted the observed count cumulatively. This statistic plotted is intuitive and it has gained widespread attention and adoption. Steiner et al. (2000) then proposed the use of a cumulative sum (CUSUM) chart that accounts for the patient's

mortality risk. It is formulated based on testing the odds ratio of the mortality. Moustakides (1986) showed that the CUSUM chart is optimal in terms of run length performance. Moreover, Rogers et al. (2004) stated that "it has the advantage of providing a formal test of an explicit hypothesis" and Spiegelhalter (2004) also mentioned that this risk-adjusted CUSUM chart "formally provides a more powerful test."

However, Rogers et al. (2004) voiced their concerns about the effect of changes in the underlying mortality risk distribution on the performance of the risk-adjusted CUSUM chart. We demonstrate this using a real data set. The data comprises the outcomes of patients with an acute myocardial infarction (more commonly known as heart attack) who are admitted to an anonymous hospital, collected as part of the NHS Research and Development funded EMMACE-1 (Evaluation of Methods and Management of Acute Coronary Events) Study (Dorsch et al. 2000). The post-operative outcomes after thirty days were collected for patients admitted over a 3-month period. The given corresponding mortality risk for each patient was both calculated and authenticated locally at the hospital. For the monitoring of the clinical performance, we adopt the risk-adjusted CUSUM charts proposed by Steiner et al. (2000) (summarized in Appendix A) while for the monitoring of the mortality risk distribution, we use Page (1954)'s CUSUM procedure (summarized in Appendix B). The CUSUM charts for this data set are shown in Figure 2.4. For the risk-adjusted CUSUM chart designed to detect improvement in performance, it signals at both 21st and 77th patients and for that designed to detect deterioration in performance, it signals 14 patients later at the 91th patient.

This leads to a suspectible conclusion that the hospital showed improvement in performance initially and yet showed a change to that of deterioration in performance over a short period. Could this conjecture be due to other reasons? The CUSUM chart in Figure 2.4(c) to detect an upward shift in the average mortality risk shows a change in pattern after the 76th patient, and it signals at the 102nd patient, thus showing an increase in the average mortality risk. As there are more patients with higher mortality risk, this results in more mortalities, thus increasing the mortality rate and reaching an erroneous impression that there is a deterioration in performance when in fact there is evidence to indicate that the performance is within expectation. As such, the deterioration is possibly due to changes in the underlying mortality risk distribution, thus showing the rationality of the earlier concerns raised by Rogers et al. (2004).

For a particular mortality risk distribution, through the adjustment for the patients' mortality risks, the risk-adjusted chart developed by Steiner et al. (2000) has accounted for the variability in the mortality risk when monitoring the clinical performance. As such, the true clinical performance is not masked. However, this adjustment for the mortality risk of patients does not account for any changes in the underlying mortality risk distribution. Assume that the mortality risk distribution be modeled by beta(α, β) which is the beta distribution, parameterized by shape parameters α and β with probability density function $f(x; \alpha, \beta) = (1-x)^{(\beta-1)} x^{\alpha-1}/B(\alpha, \beta)$, where $B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$. From the plot of the probability density functions of the monitoring statistic W_t of the risk-adjusted CUSUM charts for testing $H_0 : Q = 1$ versus $H_A : Q = 2$ given the true odds

ratio $Q = 1$, in Figure 2.1, when the risk distribution changes from beta(1,3) to beta(1,5), this will result in more patients of low risk, and a corresponding increase in the proportion of negative W_t values and a decrease in the proportion of positive W_t values. Any changes in the risk distribution will result in a change in the probability density function of the monitoring statistic and hence the performance of the risk-adjusted CUSUM chart will be affected. In summary, we found that similar to most charting procedures, despite the fact that the risk-adjusted CUSUM chart has adjusted for the patients' mortality risks, it is still sensitive to changes in the risk distribution.

In order not to wrongly assess clinical performance due to changes in the risk distribution, one should jointly monitor the clinical performances and the mortality risks. In Section 2, we further investigate the effects of changes in the risk distribution on the performances of the risk-adjusted CUSUM charts proposed by Steiner et al. (2000). We also show that through the use of simulated data sets with characteristics similar to a real data set, the joint monitoring of the clinical performances and the mortality risk is essential. In Section 3, the joint monitoring scheme for the clinical performances and the mortality risk will be explained in detail and demonstrated with a real data set. In Section 4, two real applications will be provided in health care context: monitoring of clinical procedural mortality. The conclusions and important findings will be presented in the last section.

SECTION 2. IMPORTANCE OF JOINT MONITORING

We first investigate the sensitivity of the earlier discussed CUSUM charts to changes in the risk distribution by comparing the in-control average run length (ARL), by which the ARL is defined as the expected number of patients seen until a signal is issued. The basis for determining the parameters and various aspects of the sensitivity analysis will be to consider situations which mimics that of a real data set analyzed in Section 1. This basis will ensure that our sensitivity analysis studies are befitting of real-life scenarios. Since the mortality risk is between 0 and 1 and from the previous studies of the risk distribution, the theoretical model distribution for the real data set may be modeled as beta(1,3). We consider changes in the underlying risk distribution to a beta distribution with shape parameter $\alpha = 1$ but with different values of β and then examine the effect on the in-control ARL. For detecting a deterioration in the clinical performance, we consider risk-adjusted CUSUM charts optimal in detecting Q_A =1.1, 1.2, 1.3, 1.4, 1.5, 2.0 and 3.0 where Q_A is the odds ratio considered in $H_A : Q = Q_A$, while for detecting an improvement in the clinical performance, we consider risk-adjusted CUSUM charts optimal in detecting Q_A =0.9, 0.8, 0.7, 0.6, 0.5, 0.2 and 0.1. The resulting ARL's are displayed in Table 2.1. We determine the in-control ARL to be 100 for which the underlying mortality risk distribution is beta(1,3).

We note that as β decreases below 3, the risk distribution becomes more skewed to the right, thus resulting in less low-risk patients and more high-risk patients. The in-control ARL also decreases by about 3% to 13%. To the contrary, we also note that as β increases above 3, the risk distribution becomes less skewed

to the right, thus resulting in more low-risk patients and less high-risk patients. The in-control ARL also increases by about 12% to 31%. This table shows clearly how the performances of the risk-adjusted CUSUM charts are affected by changes in the risk distribution. It is thus important to monitor clinical performances and mortality risk jointly because any inferences drawn from a risk-adjusted CUSUM chart alone should be treated with caution.

We also investigate two simulated data sets with characteristics similar to the real data set as mentioned earlier, to further illustrate the importance of simultaneous monitoring of the clinical performances and the mortality risk:
(1) A data set in which the 100 patients' risk follow the beta(1,3) distribution, with the clinical performance meeting expectation for the first 50 patients but had deteriorated (with the odds of mortality increasing by 2 fold) for the last 50 patients;
(2) A data set in which the first 50 patients' risk follow the beta(1,3) distribution and the last 50 patients' risk follow the beta(1,2.5) distribution, with the performance meeting expectation for all 100 patients.
For each data set, the risk-adjusted CUSUM charts for detecting deterioration and improvement, as well as the CUSUM charts for the monitoring of the average mortality risk, are run simultaneously. The CUSUM charts for the 2 simulated data sets are shown in Figures 2.2 and 2.3 respectively.

For the first data set, the risk-adjusted CUSUM chart in Figure 2.2(a) shows an obvious change in pattern after the 66th patient, and it signals at the 84th and 100th patients, with no changes in the risk distribution as shown by the charts in

Figures 2.2(c) and 2.2(d). This shows that there is a deterioration in performance, with no changes in the risk distribution. For the second data set, the risk-adjusted CUSUM chart in Figure 2.3(a) also shows an obvious change in pattern after the 51st patient, and it signals at the 70th and 92th patients, thus showing that there is also a deterioration in performance. But the CUSUM chart in Figure 2.3(c) to detect an upward shift in the average mortality risk shows a change in pattern after the 50th patient, and it signals at the 59th, 89th and 100th patients, thus also showing an increase in the average mortality risk of the patients. As there are more patients with higher mortality risk, this might result in more mortalities, thus increasing the mortality rate in the data set and resulting in an erroneous impression that there is a deterioration in performance. Through the two data sets discussed, the joint monitoring of the clinical performances and the mortality risk is not just necessary but also essential because any inferences drawn from a risk-adjusted CUSUM chart alone could be erroneous when the risk distribution has changed. Indeed, if joint monitoring scheme is implemented, any inferences drawn will be more indicative of the true clinical performances.

SECTION 3. DESIGN OF JOINT MONITORING SCHEME

In this section, a joint monitoring scheme for the clinical performances and the average mortality risk is described in detail. The illustration of this monitoring scheme will be based on the real data analyzed in Section 1. There are 4 steps for constructing each of the charts, whether it is to monitor either deterioration or improvement in performance, or an upward or downward shift in the average

mortality risk.

Step 1. Determine the mortality risk distribution of the patients.

Step 2. Decide on the false signal rate for the charts.

Step 3. Decide on a threshold of an unacceptable value for each parameter of interest.

Step 4. Determine the control chart parameters.

Step 1. Determine the mortality risk distribution of the patients

Before a monitoring scheme is introduced, Woodall (2000) recommended that it is evaluated using a Phase I analysis of historical data and a Phase II monitoring. Steiner (2006) and Burkom (2006) also recommended using Phase I/Phase II studies to assess any health care control charts. For the Phase I study, the risk factors present for a group of patients, as well as their post-procedural outcomes are recorded. Once sufficient data are collected in conjunction with an audit of the on-going clinical performance to ensure that the process is in-control, the mortality risks for the patients may then be determined by using a rating method, such as Parsonnet risk factors (Parsonnet, Dean and Bernstein 1989). Afterwhich, a logistic regression model is used to convert these scores obtained from the rating method, to a risk value between 0 and 1. The risk may also be computed based on a logistic regression model fitted to sample data or past data set, such as the EuroSCORE (Nashef et al., 1999) which is used to evaluate the risk of patients for cardiac operations. Based on the risks obtained in this retrospective analysis, explanatory techniques such as probability plots and histograms, can first

be employed to study the shape of the underlying risk distribution. Numerical methods, such as Kolmogorov-Smirnov test, Anderson-Darling test or chi-square goodness-of-fit test, can then be used to ascertain the risk distribution. For a beta distribution, the parameters α and β can also be easily estimated using method-of-moments estimates as

$$\hat{\alpha} = \bar{x}[\bar{x}(1-\bar{x})/s^2 - 1], \tag{2.1}$$

$$\hat{\beta} = (1-\bar{x})[\bar{x}(1-\bar{x})/s^2 - 1], \tag{2.2}$$

where $\bar{x}$ is the sample average and s^2 is the sample variance of the mortality risks obtained in a Phase I study. For the hospital in the EMMACE-1 study that we studied, the beta(1,3) distribution is found to provide an adequate fit to the data. This results in a average mortality risk of $1/(1+3)$ or 25%, and it is fairly consistent with the overall mortality rate of about 21.3%.

Step 2. Decide on the false signal rate for the charts

A false signal rate θ implies that on average, $1/\theta$ runs will be plotted until a signal is issued when the process is in control. This is equivalent to stating the in-control ARL as $1/\theta$. Suppose the average number of patients admitted per year is 800 and hospital administrators decide that 4 false signals per year is reasonable. This results in a false signal rate of 4 per 800 patients, or 1 per 200 patients to be plotted on the chart. For another scenario, if the hospital administrators decide that 8 false signals per year is reasonable, then this results in a false signal rate of 8 per 800 patients, or 1 per 100 patients to be plotted on the chart. This false

signal of 1 per 100 patients would mean that on average, out of every 100 patients admitted, the chart will issue a signal that the process might have changed even though the process is in control.

The choice of an appropriate false signal rate θ depends primarily on the desired Type I error rate. Spiegelhalter et al. (2003) stated that the desired Type I error rate should reflect the relative "costs" of making the error. For example, if we wish to avoid falsely identifying a clinical procedure is performing beyond expectations, we will select a small Type I error rate which corresponds to a small false signal rate. Although a low false signal rate is desirable, it is noted that a chart with a lower false signal rate will take longer to signal when the process has changed. This trade off should be considered carefully in the determination of an appropriate false signal rate.

The number of patients admitted is essentially important as well. Suppose the average number of patients admitted per year is 100 and hospital administrators decide that the false signal rate is 1 per 200 patients. This indicates that on average, the chart will issue a signal every 2 years even though the process is in control. The chart will also take a long time to signal when the process has changed. As such, if the number of patients admitted for the clinical procedure is low, the appropriate false signal rate will usually be pre-determined higher.

For the hospital in the EMMACE-1 study that we studied, a false signal rate of 1 per 200 patients is determined as the number of patients admitted is relatively large. Various false signal rates have also been used in practice. For example, a false signal rate of 1 per 400 patients was used in the monitoring of the occurrences

of surgical wound infections (Sherlaw-Johnson et al., 2005), and a false signal rate of 1 per 100 patients was proposed by Coory, Duckett and Sketcher-Baker (2008) in the monitoring of the quality of hospital care using administrative data.

Suppose that the in-control ARL for each of the plots are $ARL^1_+ = ARL^1_- = ARL^2_+ = ARL^2_- = 200$ where ARL^1_+ and ARL^1_- are the in-control ARLs for the charts to monitor clinical performances, and ARL^2_+ and ARL^2_- are that for the charts to monitor average mortality risk, with + referring to the monitoring an improvement in performance or upward shift in the average mortality risk and − referring to the monitoring an deterioration in performance or downward shift in the average mortality risk. The overall ARL^* can be approximated by using:

$$\frac{1}{ARL^*} \approx \frac{1}{ARL^1_+} + \frac{1}{ARL^1_-} + \frac{1}{ARL^2_+} + \frac{1}{ARL^2_-}, \tag{2.3}$$

Step 3. Decide on a threshold of an unacceptable value for each parameter of interest

For the monitoring of clinical performances, the odds ratio Q_0 in H_0 is set to be 1 which indicates that the patient care process is performing within expectations under current conditions. The odds ratio Q_A in H_A is usually taken to be the threshold of an unacceptable odds ratio for an outcome when testing for deterioration or improvement. In order to detect a deterioration, it is similar to detect an increase in the mortality rate, thus we will set $Q_A > 1$ but if the intent is to detect an improvement, it is similar to detect a decrease in the mortality rate, thus we will set $Q_A < 1$. Two different risk-adjusted CUSUM charts are required, with one

for the detection of improvement and the other for detection of deterioration. This is necessary because the monitoring statistic W_t depends on the odds ratio Q_A, which is different when testing for both improvement and deterioration. Steiner et al. (2000) proposed using odds ratio $Q_A = 2$ and 0.5 which represents halving and doubling the odds of mortality respectively. Novick et al. (2006) provided alternative values of the odds ratio, $Q_A = 3/2$ and $2/3$ for monitoring coronary artery bypass graft surgical outcomes. For monitoring mortality rates in interventional cardiology, Matheny, Ohno-Machado and Resnic (2007) used $Q_A = 3/2$ and 2 as the study is interested in monitoring whether the mortality rates have increased. For the hospital in the EMMACE-1 study that we studied, we determined the thresholds for the odds ratio to be $Q_A = 2$ and 0.5.

To monitor the average mortality risk for a beta distribution, the average mortality risk μ_0 is set to be $\bar{x}$, which is the sample average of the mortality risks obtained in the Phase I study, as discussed in Step 1. If other distributions for the mortality risk are proposed, the average mortality risk μ_0 can be taken as the average for the proposed distribution. Two different CUSUM charts are also required, with one for the detection of an upward shift in the average mortality risk and the other for the detection of a downward shift. The corresponding shifted average mortality risk μ_1 is set such that $\mu_1 > \mu_0$ and $\mu_1 < \mu_0$ respectively. This is again necessary because the score W_t, as shown in Appendix B, depends on this shifted average mortality risk μ_1, which is different when testing for both an upward shift and a downward shift. We propose to set $\mu_1 = 1.2\mu_0$ and $\mu_1 = 0.8\mu_0$ to detect an upward shift and a downward shift respectively. This will correspond

to a 20% increase and a 20% decrease in the average mortality risk respectively. For the hospital in the EMMACE-1 study that we studied, we determined the thresholds for the shifted average mortality risk to be $\mu_1 = 0.3$ and 0.2, with $\mu_0 = 0.25$.

Step 4. Determine the control chart parameters

Upon setting the false signal rate θ and the parameters of interest Q_A and μ_1 in H_A in steps 2 and 3, the control chart parameter, specifically the upper control limit for each chart can then be determined such that it produces the specified in-control ARL=$1/\theta$. To achieve this, the collocation method proposed by Knoth (2005, 2007) is used to compute the ARL for a fixed control limit of the chart. Details can be found in Appendix C. Alternatively, the control chart parameters can be determined using simulation.

For the hospital in the EMMACE-1 study that we studied, the false signal rate is set as 1 per 200 patients. We also determined that the thresholds for the odds ratio to be $Q_A = 0.5$ and 2, and that the thresholds for the shifted average mortality risk to be $\mu_1 = 0.3$ and 0.2 with $\mu_0 = 0.25$. With these values, the chart parameters can be determined. The control limit of the chart for detecting a deterioration in performance and that for detecting an improvement in performance are determined as 2.107 and 2.000 respectively. The control limit of the chart for the detection of an upward shift in the average mortality risk and that for the detection of a downward shift are determined as 7.699 and 8.733 respectively. These control limits are determined using a computer program developed by the

authors and it is available upon request.

The resulting CUSUM charts for this example are shown in Figure 2.4. As mentioned in the introduction, for the risk-adjusted CUSUM chart designed to detect improvement in performance, it signals at both 21st and 77th patients but for that designed to detect deterioration, it signals 14 patients later at the 91th patient. This leads to a suspectible conclusion that the hospital showed improvement in performance initially and yet showed deterioration over a short period. The CUSUM chart in Figure 2.4(c) to detect an upward shift in the average mortality risk shows a change in pattern after the 76th patient, and it signals at the 102nd patient, thus showing an increase in the average mortality risk. As there are more patients with higher risk, this results in more mortalities, thus increasing the mortality rate and resulting in an erroneous impression that there is a deterioration in performance when there is evidence to indicate that the performance is within expectation.

SECTION 4. REAL APPLICATIONS

To better reiterate our proposed charting procedures, illustrations for two real applications are shown. The two real data sets are obtained from an anonymous hospital in UK. For this data set, the patients underwent two different type of cardiac surgery operations in the hospital and their post-operative outcomes after thirty days were collected. The corresponding mortality risk x_t for each patient was both calculated and authenticated locally at the hospital.

For the first example, a Phase I analysis of historical data with an audit of

the clinical performance is conducted. This is to ensure that the Phase I analysis is conducted using data in which the clinical performance is in-control. A total of 71 patients over a period of time are considered. Using the method-of-moments estimates in (2.1) and (2.2), $\hat{\alpha} = 5.162$ and $\hat{\beta} = 24.337$. Due to low admission rate for this type of cardiac surgery operation, the false signal rate is determined as 1 per 50 patients. We also determine the thresholds for the odds ratio to be $Q_A = 2$ and 0.5, and that for the shifted average mortality risk to be $\mu_1 = 0.210$ and 0.140 with $\mu_0 = 0.175$. With these information, the control limits of the chart for detecting deterioration in performance and that for detecting improvement in performance are determined as 1.184 and 1.072 respectively. The control limits of the chart for the detection of an upward shift in the average mortality risk and that for the detection of a downward shift are also determined as 1.317 and 1.419 respectively.

The Phase II monitoring is conducted subsequently for 67 patients and the CUSUM charts for this example are shown in Figure 2.5. For the risk-adjusted CUSUM chart designed to detect a deterioration in performance, it shows a change in pattern after the 22nd patient, and it signals at the 39th patient, but for that designed to detect an improvement in performance, it signals at the 64th patient. This again leads to a suspectible conclusion that the hospital showed a deterioration in performance initially and thereafter showed an improvement in performance. The CUSUM chart in Figure 2.5(d) to detect a downward shift in the average mortality risk shows a signal at the 63rd patient, thus showing a decrease in the average mortality risk of the patients. Due to more patients with lower mor-

tality risk, this might result in less mortalities, thus decreasing the mortality rate and resulting in an erroneous impression that there might be an improvement in performance. We can only conclude there is evidence that the hospital experiences a deterioration in performance.

For the next example, another Phase I analysis of historical data with an audit of the clinical performance is conducted. A total of 71 patients over a period of time are considered. Using the method-of-moments estimates in (2.1) and (2.2), $\hat{\alpha} = 1.093$ and $\hat{\beta} = 6.772$. Due to low admission rate for this type of cardiac surgery operation again, the false signal rate is determined as 1 per 50 patients. We determine the thresholds for the odds ratio to be $Q_A = 2$ and 0.5, and that for the shifted average mortality risk to be $\mu_1 = 0.167$ and 0.111 with $\mu_0 = 0.139$. With these information, the control limits of the chart for detecting deterioration in performance and that for detecting improvement in performance are determined as 1.045 and 0.934 respectively. The control limits of the chart for the detection of an upward shift in the average mortality risk and that for the detection of a downward shift are also determined as 4.460 and 4.793 respectively.

The Phase II monitoring is conducted subsequently for 54 patients and the CUSUM charts for this example are shown in Figure 2.6. For the risk-adjusted CUSUM chart designed to detect an improvement in performance, it signals at the 13th and 26th patients, while the CUSUM chart designed to detect a deterioration in performance signals at the 50th patient. We are led to a suspectible conclusion that the hospital showed an improvement in performance initially and thereafter showed a deterioration in performance. Looking at the CUSUM charts in Figure

2.6(c) and 2.6(d), the CUSUM chart in Figure 2.6(d) to detect an upward shift in the average mortality risk signals at the 18th and 30th patient, thus showing an increase in the average mortality risk of the patients before the 30th patient. This indicates there is evidence that the hospital is not just showing an improvement in performance, it is in fact showing exemplary performance in reducing the odds of mortality despite experiencing an increase in the average mortality risk of the patients. However, the hospital also subsequently shows a deterioration in performance because there is no evidence of any change in the average mortality risk.

SECTION 5. CONCLUSIONS

Measuring quality of medical practice is a key component in improving efficiency in health care, such assessment is playing an increasingly prominent role in quality management. One fundamental practice of assessment will be that of clinical performance monitoring. In this paper, we introduce a new charting procedure to monitor the mortality risk distribution, specifically the average mortality risk of patients. Although the proposed procedure is used to monitor the average mortality risk, with the risk modeled by a beta distribution, through slight modifications, this charting procedure can be used for other distributions for the risk.

More importantly, we propose to jointly monitor the clinical performances and the mortality risk. Rogers et al. (2004) expressed their concerns about the effect of changes in the underlying mortality risk distribution on the performance of the risk-adjusted CUSUM charts used to monitor clinical performances. By using

a sensitivity analysis study of the effects of changes in the risk distribution on the in-control ARL, as well as illustrations using real applications and simulated examples, our findings suggest that any inferences drawn from a risk-adjusted CUSUM chart alone could be erroneous when the risk distribution has changed. Indeed, if joint monitoring scheme is implemented, any inferences drawn will be more indicative of the true clinical performances. The monitoring of the mortality risk provides a better understanding for any inferences drawn from the risk-adjusted CUSUM charts. In fact, the joint monitoring of the clinical performances and the mortality risk is not just necessary but also essential.

The design of the joint monitoring scheme for the clinical performances and the average mortality risk is also described in detail, with an illustration based on a real data set. It is important to note that the implementation of the joint monitoring scheme is able to adequately identify probable changes in the clinical performances and mortality risk distribution, controlling for all possible risk-adjusting factors. Only upon seeking out these probable changes, there can begin a process to further improve the performances, which may include retraining of staff or upgrading of equipment. As such, we urge that joint monitoring of the clinical performances and the mortality risk needs to become an integral part in the measurement of the quality of medical practice.

Acknowledgement

We wish to thank Dr Alistair Hall for providing the data from the EMMACE-1 Study and the permission to use it here.

Table 2.1. In-control average run lengths of risk-adjusted CUSUM charts based on testing odds ratio corresponding to various underlying risk distributions

Risk Distribution	$Q_A = 1.1$ $h = 0.308$	$Q_A = 1.2$ $h = 0.558$	$Q_A = 1.3$ $h = 0.765$	$Q_A = 1.4$ $h = 0.940$	$Q_A = 1.5$ $h = 1.09$	$Q_A = 2.0$ $h = 1.607$	$Q_A = 3.0$ $h = 2.125$
beta(1,2)	91	91	92	92	92	94	95
beta(1,2.5)	95	95	95	96	96	96	97
beta(1,3)	100	100	100	100	100	100	100
beta(1,4)	112	111	111	110	110	109	107
beta(1,5)	124	123	123	122	121	119	116

Risk Distribution	$Q_A = 0.9$ $h = 0.335$	$Q_A = 0.8$ $h = 0.652$	$Q_A = 0.7$ $h = 0.954$	$Q_A = 0.6$ $h = 1.242$	$Q_A = 0.5$ $h = 1.521$	$Q_A = 0.2$ $h = 2.330$	$Q_A = 0.1$ $h = 2.616$
beta(1,2)	90	90	90	89	89	87	87
beta(1,2.5)	95	95	94	94	94	93	93
beta(1,3)	100	100	100	100	100	100	100
beta(1,4)	112	112	113	113	113	115	115
beta(1,5)	125	126	126	127	127	130	131

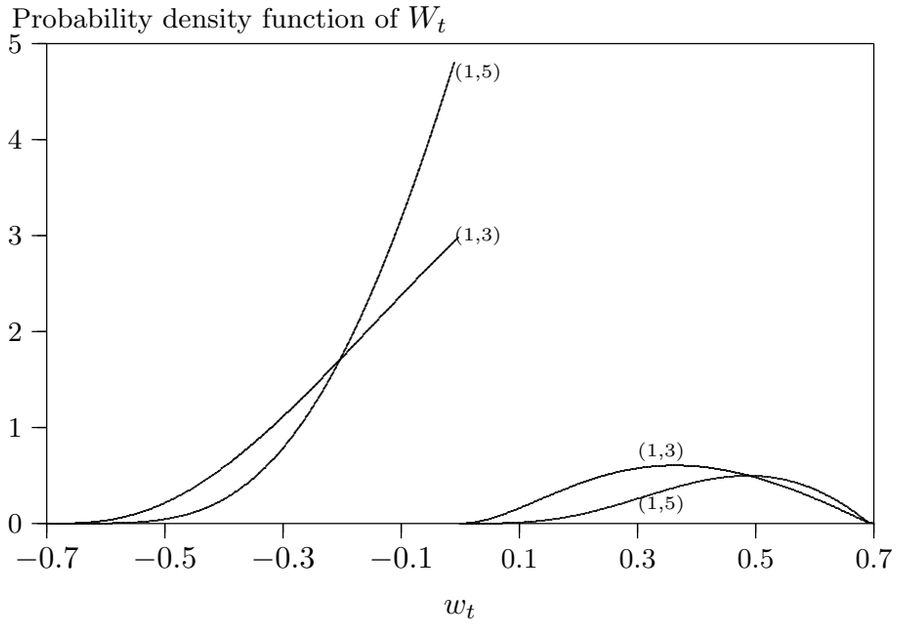

Figure 2.1. Probability density functions of the monitoring statistic W_t of the risk-adjusted CUSUM chart proposed by Steiner et al. (2000) for testing $H_0 : Q = 1$ versus $H_A : Q = 2$ given the true odds ratio $Q = 1$, corresponding to mortality risk distributions beta(1, 3) and beta(1, 5).

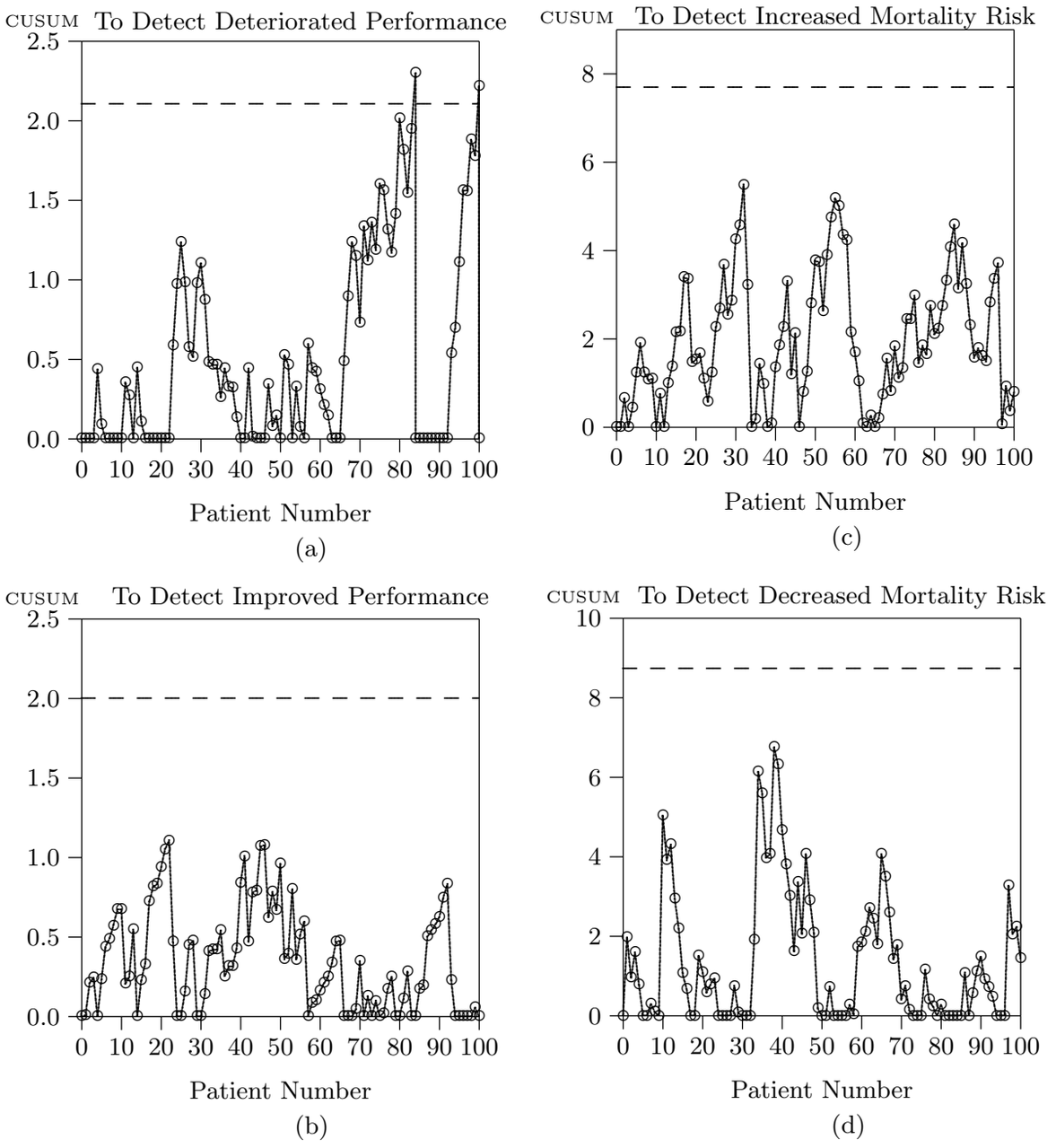

Figure 2.2. CUSUM charts to detect (a) deterioration in performance, (b) improvement in performance, (c) upward shift in the average mortality risk and (d) downward shift in the average mortality risk, for a data set in which the 100 patients' risk follow the beta(1,3) distribution, with the performance meeting expectation for the first 50 patients but had deteriorated for the last 50 patients. The dashed lines represent the control limits. These charts signal correctly for the deterioration in performance, with no changes in the mortality risk distribution.

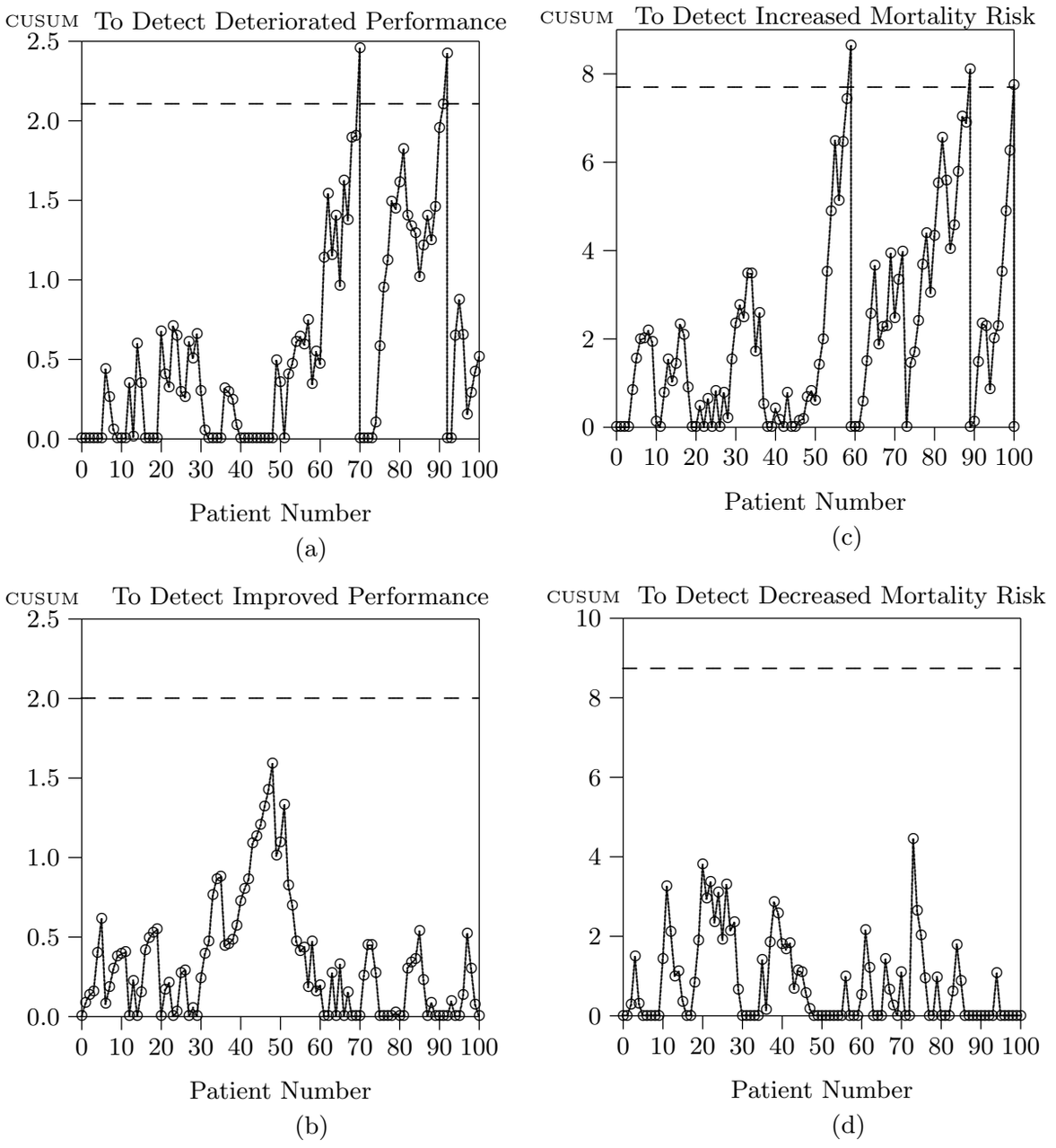

Figure 2.3. CUSUM charts to detect (a) deterioration in performance, (b) improvement in performance, (c) upward shift in the average mortality risk and (d) downward shift in the average mortality risk, for a data set in which the first 50 patients' risk follow the beta(1,3) distribution and the last 50 patients' risk follow the beta(1,2.5) distribution, with the performance meeting expectation for all 100 patients. The dashed lines represent the control limits. These charts signal incorrectly for the deterioration in performance when in fact the signal is due to the higher risks of the last 50 patients.

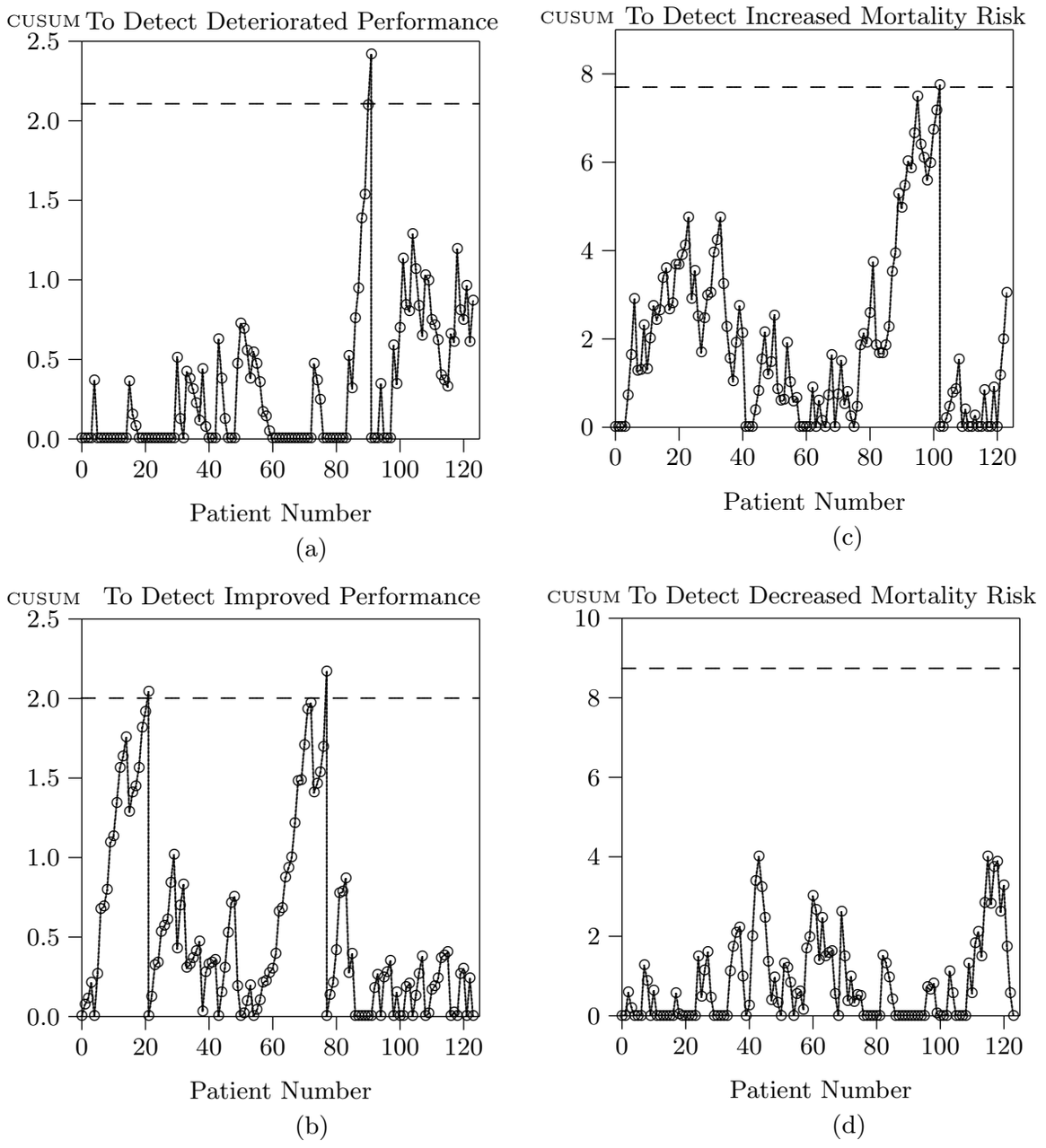

Figure 2.4. CUSUM charts to detect (a) deterioration in performance, (b) improvement in performance, (c) upward shift in the average mortality risk and (d) downward shift in the average mortality risk, for patients with an acute myocardial infarction who are admitted to an anonymous hospital, collected as part of the EMMACE-1 Study. The dashed lines represent the control limits. These charts signal for an improvement in performance initially (see (b)) and a subsequent deterioration in performance (see (a)), with the latter corresponding to an increase in the average mortality risk (see (c)). Without charts (c) and (d), one might make an erroneous conclusion that there is a deterioration in performance when there is evidence to indicate that the performance is within expectation.

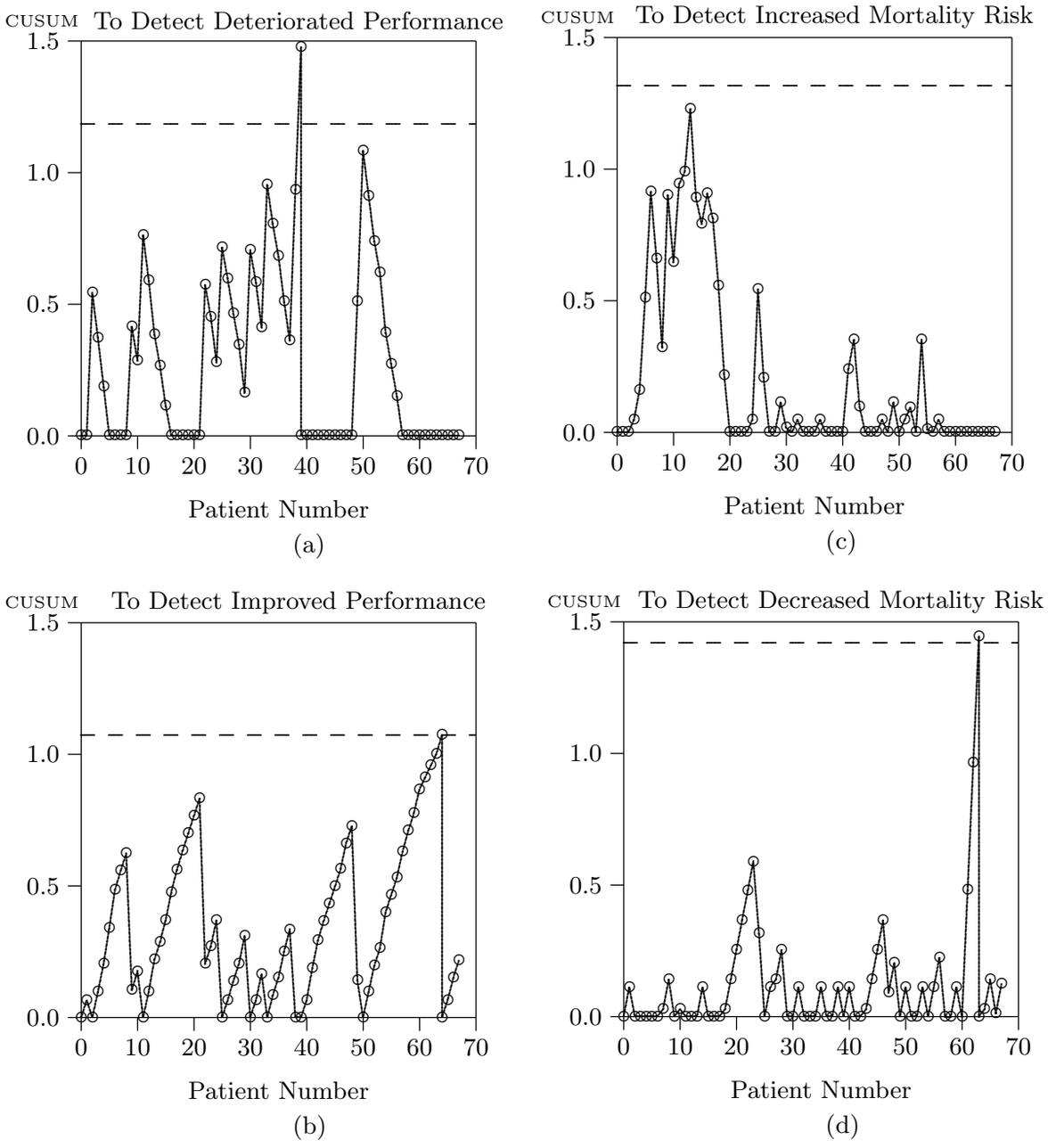

Figure 2.5. CUSUM charts to detect (a) deterioration in performance, (b) improvement in performance, (c) a upward shift in the average mortality risk and (d) downward shift in the average mortality risk, for patients who underwent cardiac surgeries in an anonymous hospital in UK. The dashed lines represent the control limits. These charts signal for a deterioration in performance initially (see (a)) and a subsequent improvement in performance (see (b)), with the latter corresponding to a decrease in the average mortality risk (see (d)). Without charts (c) and (d), one might make an erroneous conclusion that there is an improvement in performance when there is evidence to indicate that the performance might be within expectation.

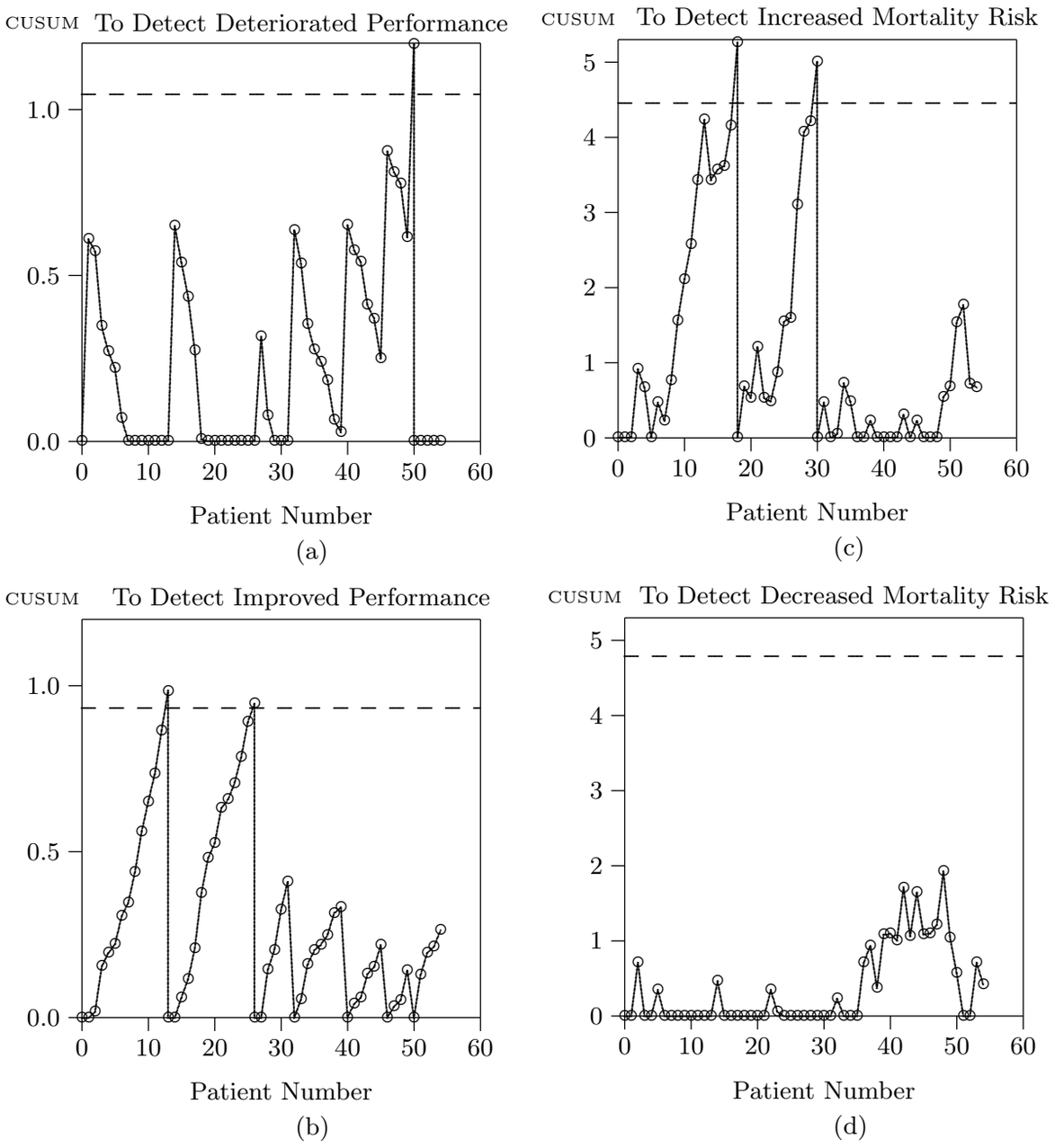

Figure 2.6. CUSUM charts to detect (a) deterioration in performance, (b) improvement in performance, (c) a upward shift in the average mortality risk and (d) downward shift in the average mortality risk, for patients who underwent cardiac surgeries in an anonymous hospital in UK. The dashed lines represent the control limits. These charts signal for an improvement in performance (see (b)) with an increase in the average mortality risk initially (see (c)), and a subsequent deterioration in performance (see (a)). There is evidence that the hospital is showing exemplary performance despite experiencing an increase in the average mortality risk. However, the hospital also subsequently shows a deterioration in performance, with no evidence of change in the average mortality risk.

CHAPTER 3: DIAGNOSTIC TECHNIQUES FOR INVESTIGATING MORTALITY RATES AND RISK-ADJUSTED METHODS FOR COMPARING TWO OR MORE CLINICAL PROCEDURES WITH VARIABLE DEGREE IN PERFORMANCE DIFFERENCES ACROSS MORTALITY RISKS

SUMMARY

The evolution of the assessment of medical practice has been speeding up tremendously. At present, risk-adjusted analytical tools are best used as a monitoring procedure, rather than to compare clinical performances. In this paper, we propose a model-free diagnostic technique to estimate the actual mortality rates for all levels of predicted mortality risk to assess clinical performances. Using the estimated mortality rates, we present a set of risk-adjusted test procedures which alleviate the problem of interpretation through the use of penalty-reward scores. We also consider other risk-adjusted methods for this comparison. Using real data, we show how the proposed diagnostic technique and various hypothesis test procedures can be used effectively to evaluate the performances of two clinical procedures. A simulation study is also conducted to investigate the performances of the proposed test procedures against a popularly-used method, the McNemar's test of equality of paired proportions.

SECTION 1. INTRODUCTION

The evolution of the assessment of medical practice has been speeding up tremendously, as seen from recent literature (Werner and Bradlow, 2006, Clarke and Oakley, 2007, Krumholz et al., 2008). Measuring quality of medical practice is a key component in improving efficiency in health care, such assessment is playing an increasingly prominent role in quality management. In recent years, the United States Centers for Medicare and Medicaid Services has been collaborating with various health care organizations to participate in the Hospital Quality Alliance (2006) such that performance information are made readily accessible to the public, payers and providers of care. It is therefore crucial that information released is reasonably accurate and fairly representative such that it is of significant value.

But the release of such performance report cards might lead to misinterpretation of the data. Patients in hospitals tend to differ notably in terms of preoperative risk of procedural failure, which in this paper we will refer to as mortality risk. If this variability in the mortality risk is not taken into account when assessing a particular physician's performance or effectiveness of a certain clinical procedure, this variability might result in additional fluctuation in the outcomes, thus masking the effect of the true performance of the physician or effectiveness of the clinical procedure, and resulting in misapprehension of the true situation. For example, if a particular physician or clinical procedure has a relatively low mortality rate, it will give an impression that this physician is highly skilled or this clinical procedure is effective, and vice versa. As such, the New York State

Department of Health (2008) do not just publish raw mortality rates, they also publish "risk-adjusted" mortality rates, which is an indication of what a physician's mortality rate would have been, had he or she treated patients identical to the state's average.

To ensure that such mortality risks are taken into account, McNemar's (1947) test of equality of paired proportions is usually employed. For example, Chen, Connors and Garland (2008) studied 201 patients, matching each of these 201 patients to another patient having the closest propensity score. The propensity score in this study is the probability of a patient of having an order initiated in the ICU to withhold life-supporting therapies. This matching procedure resulted in the matched pairs being well-balanced with respect to all the potentially confounding variables. Some other reported applications of the McNemar's test include Maxwell (1970), Cardozo et al. (1980), Altman et al. (1983), Seeman et al. (1983), Schatzkin et al. (1987), Uhlmann, Pearhman and Cain (1988), Schwartz et al. (1991), Greinacher et al. (1994), Johnston et al. (1995), Egger et al. (1997), Kuipers et al. (1996), Scott, Besag and Neville (1999), Dickerson et al. (1999), Dooley et al. (2001), Koopmans et al. (2008), Berger et al. (2008), Quigley et al. (2008), Yan et al. (2008) and Boccasanta et al. (2009). It is interesting to note that the McNemar's test only focuses on the matched pairs in which the outcomes differ for the members of the pairs, more commonly known as the discordant pairs. This indicates that the matched pairs in which the outcome is the same for each member of the pairs, or the concordant pairs, are not utilized in the assessment, thus possibly losing valuable information from the data. Moreover, it is also im-

portant to note that in order for the matching procedure to be conducted, it is inefficient through the implementation of such a procedure. For example, in Chen, Connors and Garland (2008), the initial number of patients observed was 2211. But in order to achieve balanced groups for the comparison, only 402 patients were studied. This implied that information from more than 80% of the initially observed patients was not taken into account. The loss in information also results in a loss in power of the McNemar's test. This is shown by the results in Table 3.1 obtained from our simulation study which is befitting of a real-life scenario.

Although the use of matched pairs will take into account of the mortality risks and thus resulting in well-balanced pairs with respect to all the potentially confounding variables, it is assumed that the effect of the true performance of the clinical procedure is the same (that is, the degree of the differences between each pair is the same) regardless of the mortality risks. But this assumption does not always hold. For example, it is possible that a certain clinical procedure works well on patients of lower risk but might not be as effective on patients of higher risk. One such scenario will be present in the treatment of coronary heart disease. Coronary angioplasty is the therapeutic procedure to treat coronary arteries of the heart that are narrowed. It is accomplished by inserting a small balloon catheter into an artery in the groin or arm, and this catheter is subsequently advanced to the narrowing in the affected narrowed coronary artery. This surgical procedure is recommended for patients of lower risk and might not be as effective for patients of higher risk, such as patients with diabetes or patients with multiple narrowings in multiple coronary arteries. Another surgical procedure, coronary

artery bypass graft (CABG) surgery is usually conducted on patients of higher risk. This procedure creates new routes around narrowed and blocked arteries, thus promoting blood flow. But due to the nature of this procedure, such an open surgery increases the amount of risks and complications faced by the patients, thus it is not recommended for patients of lower risk. Consequently, it is important to note that current literature of the test procedures examine the hypothesis that assumes the degree of the differences between the two clinical procedures is the same, such as the McNemar's test. If the effect of the true performance of the clinical procedure is different across the range of mortality risks, the power of these test procedures will be greatly undermined. In fact, these test procedures are inappropriate to be applied under such scenarios.

In Section 2, we will examine the use of logistic regression to compare performances of clinical procedures. However, logistics models are usually set up by assuming a linear relationship between the logistic function of mortality rates, and mortality risks, which in the event of a wrong assumed model, the power of the test will be diminished. Unlike linear regression, there is no R^2 associated with a logistic model, thus it is not simple to evaluate whether a model is wrongly used. As such, in this section, we will also propose a model-free diagnostic technique to evaluate the effectiveness of the clinical procedure by investigating the mortality rates and resulting odds ratio function against the mortality risks. Inspired by Steiner et al. (2000), we will then proceed to formulate test procedures by making modifications to the usual logistic model. We will also show that the log-likelihood ratio scores for a patient proposed proposed by Steiner et al. (2000) can be inter-

preted as a penalty-reward score given to a particular clinical procedure and this will be used to formulate an alternative risk-adjusted procedures for comparing two or more clinical procedures. In Section 3, two real examples will be provided in health care context: clinical procedural mortality. Using a simulation study, the comparison of the proposed and McNemar's test procedures will also be analyzed and their corresponding efficiencies will be presented in Section 4. This will also allow us to illustrate the advantages of using the proposed risk-adjusted test procedures over the McNemar's test procedures. The conclusions and important findings will then be presented in the last section.

SECTION 2. GENERAL APPROACH FOR PROPOSED RISK-ADJUSTED PROCEDURE

Monitoring of the effectiveness of clinical procedures and physicians' performance has been popularized well over 40 years ago in the medical field (Armitage, 1954 and Bartholomay, 1957) but it was only till 1997 when Lovegrove et al. (1997, 1999) and Poloniecki, Valencia and Littlejohns (1998) independently developed the variable life-adjusted display (VLAD) and cumulative risk-adjusted mortality (CRAM) charts respectively, in which the mortality risk of patients is taken into account. For the health care delivery, patients in hospitals will differ notably in terms of mortality risk. An adjustment for prior risk has to be implemented to ensure that the effect of the true performance of the clinical procedure is not masked by the variability in this prior risk.

If we let y to be the post-procedural outcome for a patient, it corresponds to one of two possible outcomes (success or failure). We assume y_t is the outcome for

patient t. (1 if there is a mortality or 0 if a patient survives after implementation of the clinical procedure). Notice that the outcome of the clinical procedure may not be observed immediately after its implementation, with one example being that for cardiac operations in which the outcome of mortality is usually determined within 30 days from surgery. If patient t dies anytime within 30 days from the surgery, y_t will be assigned a value of 1 and if the patient survives after 30 days from surgery, y_t will be assigned a value of 0. As a result, we have the following probability function of y_t, $f(y_t|p) = p^{y_t}[1-p]^{1-y_t}$, where p is the mortality rate.

We further assume x_t to be the mortality risk for patient t and it is estimated prior to the implementation of the clinical procedure and it depends on the risk factors present for the patient. This risk can be determined by using a rating method, such as Parsonnet risk factors (Parsonnet, Dean and Bernstein, 1989) for cardiac operations. Afterwhich, a logistic regression model is used to convert these scores obtained from the rating method, to a risk value between 0 and 1. The risk may also be computed based on a logistic regression model fitted to sample data or past data set, such as the EuroSCORE (Nashef et al., 1999) which is used to evaluate the risk of patients for cardiac operations.

Section 2.1 Usual Logistic Regression Test Procedures for Comparing Clinical Procedures

For the health care delivery, the importance to monitor the effectiveness of clinical procedures has also been discussed, as seen from well-publicized cases (Werner and Bradlow, 2006, Clarke and Oakley, 2007, Krumholz et al., 2008). This will allow providers of care to investigate if there is a need for procedural changes

promptly. For the generality of this paper, the emphasis will be placed on the discussion of the comparison between the effectiveness of two clinical procedures.

Suppose that we are interested in comparing the performance of two clinical procedures: Procedure 1 and Procedure 2, assuming that their performances are not affected, inter alia by other environmental factors. A collection of samples of the patients treated upon using each Procedure i ($i = 1, 2$) has been collected. Also suppose that this set of n_1 and n_2 samples of bivariate data (x_{it}, y_{it}) has been collected for Procedures 1 and 2 respectively, in which we observe the patient's mortality risk and post-procedural outcome.

In order to compare the performance of two clinical procedures, one is testing the hypotheses, H_0 : Performances of both procedures are the same versus H_A : Performances of both procedures are different. Intuitively, due to the data setting of a dichotomous categorical dependent variable y_{it} and a predictor variable x_{it}, we can utilize the usual logistic model $logit[p(x_{it})] = \beta_0 + \beta_1 x_{it}$. In order to compare the performance of two clinical procedures, one can consider comparing the following logistic models:

$$logit[p(x_{it})] = \beta_0 + \beta_1 x_{it}, \tag{3.1}$$

versus

$$logit[p(x_{it})] = \beta_0 + \beta_1 x_{it} + \beta_2 I(i = 1) + \beta_3 x_{it} I(i = 1), \tag{3.2}$$

where $I(i = 1)$ is an indicator function with a value 1 if we are appraising Procedure 1, or 0 if we are appraising Procedure 2. If there is no difference between the performances of the two procedures, this will result in (3.1) and (3.2) to be the

same model, with β_2 and β_3 both being equal to 0. But if their performances are indeed different, the pairs of logistic models will be different.

In furtherance of this comparison, one can utilize the deviance goodness-of-fit test (McCullagh and Nelder, 1989), where (3.2) is the full (or saturated) model, and (3.1) is the reduced model respectively. The deviance statistic for the full model will be subtracted from the corresponding deviance statistic for the reduced model, where the deviance statistic for one model is:

$$D(\mathbf{p}(\mathbf{x_{it}}); \beta) = -2\Big[\ell(\beta; \mathbf{p}(\mathbf{x_{it}})) - \ell(\beta_{\mathbf{max}}; \mathbf{p}(\mathbf{x_{it}}))\Big] \tag{3.3}$$

where $\ell(\beta; \mathbf{W})$ is the log-likelihood statistic of that model and $\ell(\beta_{\mathbf{max}}; \mathbf{W})$ is the log-likelihood statistic of a model with a parameter β_i for every observation such that the data is fitted exactly. The difference between the residual deviance for (3.2) and (3.1) will then be tested using a χ^2-distribution with the degrees of freedom as the number of additional parameters in the full model, which is 2 in our comparison.

It is noted from the literature that it is common that the logistic models will utilize a linear function of the independent variable x_{it}. As such, this method is dependent on the model or relationship between $p(x_{it})$ and x_{it}. If a wrong model is used, this will in turn result in a less powerful test, as shown by the simulation studies in Section 4. Although this method is easily implemented using statistical software packages, it is not as easy and straightforward to test for the goodness-of-fit of each logistic models. Unlike linear regression, there is no R^2 associated with a logistic model, since residuals do not exist. As such, it is not simple to evaluate

whether a model is wrongly used. Before a proper implementation of the logistic regression procedure can be made, we will need to investigate the relationship between $p(x_{it})$ and x_{it}. In the next section, we develop a model-free diagnostic technique for this purpose.

Section 2.2 Model-Free Diagnostic Technique to Investigate Mortality Rates

Steiner et al. (2000) proposed the use of a risk-adjusted cumulative sum (CUSUM) chart that accounts for the patient's mortality risk. This risk-adjusted CUSUM chart is formulated based on testing the odds ratio of the mortality of a patient, where H_0 : odds ratio $= Q_0$ versus H_A : odds ratio $= Q_A$. This is equivalent to testing $H_0 : p_0(x_t)/[1 - p_0(x_t)] = Q_0 x_t/(1 - x_t)$ versus $H_A : p_A(x_t)/[1 - p_A(x_t)] = Q_A x_t/(1 - x_t)$ with x_t being the mortality risk for patient t and the mortality rates $p_0(x_t)$ and $p_A(x_t)$ are functions of the mortality risk x_t.

It is also noted that a scatter plot of the data $(x_{it},\ y_{it})$ will not be too informative because of the Bernoulli nature of the outcome y_{it}, other than to indicate that mortality rate may appear to increase with the mortality risk. It will be more elucidative if we are able to plot $p(x_{it})$ against x_{it}, which can subsequently be transformed to map $\{p(x_{it})[1-x_{it}]\}/\{x_{it}[1-p(x_{it})]\}$ or $Q(x_{it})$ against x_{it} such that we are able to visualize the form of the odds ratio. We will be able to investigate if the odds ratio is a constant or a function of x_{it}. In order to achieve this using a model-free approach, we will implement a two-step procedure. We will first obtain a kernel-based matching estimator $\hat{p}(x_{it}; h)$ to estimate $p(x_{it})$, and by using the plot of $\{\hat{p}(x_{it}; h)[1-x_{it}]\}/\{x_{it}[1-\hat{p}(x_{it}; h)]\}$ or $\hat{Q}(x_{it}; h)$ against x_{it}, we

will be able to identify the form of the odds ratio $Q(x_{it})$. This will allow us to identify whether the odds ratio $Q(x_{it})$ is a constant or a function of x_{it}. It is also important to note that this step does not assume any relationship between $p(x_{it})$ and x_{it}, and is model-free.

The next step will be to obtain a smoother estimate of the odds ratio function through the use of the mean square error (MSE) criterion upon the establishment of the form. We can then obtain $\hat{p}(x_{it})$ by using the smoother estimate of the odds ratio function.

For the initial step, we will need to apply an algorithm that employs a "distance" threshold to estimate $p(x_{it})$ for each x_{it}. By using kernel-based matching estimators which are commonly used in topological studies, we will form weighted averages of the post-procedural outcome y_{it} of all n patients in the sample:

$$\hat{p}(x_{it}; h) = \frac{\sum_{j=1}^{n} K(\frac{x_{it}-x_{ij}}{h}) y_{ij}}{\sum_{j=1}^{n} K(\frac{x_{it}-x_{ij}}{h})}, \tag{3.4}$$

where $K(\cdot)$ is the kernel function which is a probability density function that is symmetric about the origin and integrates to 1 over the domain, and h is a bandwidth parameter which controls the amount of smoothing of the data to obtain the estimate. We have investigated various kernel function $K(\cdot)$ developed in the literature and the Gaussian kernel function with bandwidth $h = 0.9n^{-1/5} \min\{s, IQR/2.68\}$ where IQR is the sample interquartile range and s is the sample standard deviation, proposed by Chen and Kelton (2006) provides satisfactory smoothing performance and emanates $\hat{Q}$ adequately. Details can be found in the Appendix D.

After we have estimated $p(x_{it};h)$, we will be able to obtain a more elucidative plot of $\{\hat{p}(x_{it};h)[1-x_{it}]\}/\{x_{it}[1-\hat{p}(x_{it};h)]\}$ or $\hat{Q}(x_{it};h)$ against x_{it} as this enables us to visualize the form of the odds ratio. This will allow us to differentiate whether the odds ratio $Q(x_{it})$ is a fixed constant or a function of x_{it}, such as $Q(x_{it}) = \beta_0 + \beta_1\ x_{it}$ or $Q(x_{it}) = \beta_0 + \beta_1\ x_{it} + \beta_2\ x_{it}^2$. This will allow us to verify if the assumption that the effect of the true performance of the clinical procedure is the same (on the odds ratio) regardless of the mortality risks, holds. Afterwhich, we will use the MSE:

$$MSE = E\left[\hat{p}(x_{it};h) - p(x_{it})\right]^2 = E\left[\hat{p}(x_{it};h) - \frac{Q(x_{it})x_{it}}{1 - x_{it} + Q(x_{it})x_{it}}\right]^2, \qquad (3.5)$$

as a criterion to find an estimate of the odds ratio $Q(x_{it})$. This is achieved by minimizing

$$\sum_{j=1}^{n}\left[\hat{p}(x_{ij};h) - \frac{\hat{Q}(x_{ij})x_{ij}}{1 - x_{ij} + \hat{Q}(x_{ij})x_{ij}}\right]^2, \qquad (3.6)$$

with respect to the odds ratio function $\hat{Q}(\cdot)$ and we will obtain a smoother estimate $\hat{Q}(x_{it})$ of the odds ratio function. As mentioned earlier, if need be, we can also obtain $\hat{p}(x_{it})$ by using $\hat{Q}(x_{it})$, and the plot of $\hat{p}(x_{it})$ against the mortality risk x_{it} to better visualize the effectiveness of the clinical procedure.

Alternatively, after one has identified the form of the odds ratio, despite the nonlinear relationship between $p(x_{it})$ and x_{it}, it is possible to transform x_{it} so that the substantive relationship remains nonlinear but the form of the relationship is linear in terms of its parameters (Berry and Feldman 1985) and utilize the logistic regression to visualize the effectiveness of the clinical procedure.

To further elaborate, as discussed earlier, in order to compare the performance

of two clinical procedures, one is testing the hypotheses, H_0 : Performances of both procedures are the same versus H_A : Performances of both procedures are different. It is equivalent to test $H_0 : Q_1(x_t) = Q_2(x_t)$ versus $H_A : Q_1(x_t) \neq Q_2(x_t)$, where $Q_i(x_t)$ is the odds ratio function of x_t for Procedure i. Under H_0 in which there is same performance between the clinical procedures, we note that $Q_1(x_t) = Q_2(x_t) = Q(x_t)$ and we can estimate $Q(x_t)$ by first pooling both collections of samples of the patients treated upon using each Procedure i $(i = 1, 2)$ and identifying the form of $Q(x_t)$ through the use of the kernel-based matching estimators in (3.4) with the Gaussian kernel function with the bandwidth parameter proposed by Chen and Kelton (2006), and the plot of odds ratio against x_t to determine the form of $Q(x_t)$. Finally (3.5) will be evaluated to obtain $\hat{Q}(x_t)$.

Section 2.3 Test Procedures Formulated from Logistic Regression with Knowledge of the form of $Q(x_t)$

Intuitively, from the definition of odds ratio of the mortality of a patient, we obtained $logit[p(x_t)] = log[Q(x_t)] + logit(x_t)$. Upon the identification of the form of $Q(x_t)$ as discussed earlier, if the odds ratio $Q(x_t)$ is a constant, one can consider comparing the following logistic models:

$$logit[p(x_{it})] = \beta_0 + \beta_1 logit(x_{it}), \tag{3.7}$$

versus

$$logit[p(x_{it})] = \beta_0 + \beta_1 logit(x_{it}) + \beta_2 I(i = 1), \tag{3.8}$$

or if the odds ratio $Q(x_t)$ is a function of x_t, by using Taylor series, one can

consider comparing the following logistic models:

$$logit[p(x_{it})] = \beta_0 + \beta_1 x_{it} + \beta_2 x_{it}^2 + \beta_3 logit(x_{it}), \tag{3.9}$$

versus

$$\begin{aligned} logit[p(x_{it})] = \beta_0 + \beta_1 x_{it} + \beta_2 x_{it}^2 + \beta_3 I(i=1) \\ + \beta_4 x_{it} I(i=1) + \beta_5 x_{it}^2 I(i=1) + \beta_6 logit(x_{it}), \end{aligned} \tag{3.10}$$

where $I(i=1)$ is an indicator function with a value 1 if we are appraising Procedure 1, or 0 if we are appraising Procedure 2. If there is no difference between the performances of the two procedures, depending on which pairs of logistic models are used, this will result in (3.7) and (3.8) to be the same model, with β_2 being equal to 0, or in (3.9) and (3.10) to be the same model, with β_3, β_4 and β_5 all being equal to 0. But if their performances are indeed different, the pairs of logistic models will be different.

Suppose the relationship between $logit[p(x_{it})]$ and x_{it} is nonlinear using the model-free diagnostic technique proposed in the earlier section, it is possible to transform x_{it} so that the substantive relationship remains nonlinear but the form of the relationship is linear in terms of its parameters (Berry and Feldman 1985), thus the independent variable $logit(x_{it})$ is introduced in the earlier logistic models, (3.7) to (3.10).

In furtherance of the comparison, one can utilize the deviance goodness-of-fit test (McCullagh and Nelder, 1989), where (3.8) or (3.10) is the full (or saturated) model, and (3.7) or (3.9) is the reduced model respectively, discussed earlier. The difference between the residual deviance for (3.8) and (3.7), or that for (3.10) and

(3.9) will then be tested using a χ^2-distribution with the degrees of freedom as the number of additional parameters in the full model, which is 1 or 3 respectively.

Section 2.4 Test Procedures formulated from SPRT

Investigating the risk-adjusted cumulative sum (CUSUM) chart proposed by Steiner et al. (2000), this chart is formulated based on testing the odds ratio of the mortality of a patient, where $H_0 : p_0(x_t)/[1 - p_0(x_t)] = Q_0 x_t/(1 - x_t)$ versus $H_A : p_A(x_t)/[1 - p_A(x_t)] = Q_A x_t/(1 - x_t)$. Usually $Q_0 = 1$, as the estimated risk x_t is based on the current conditions before taking into account the effect of the true performance of the clinical procedure. In order to detect an increase in the mortality rate, we will set $Q_A > Q_0$ but if the intent is to detect a decrease in the mortality rate, we will set $Q_A < Q_0$. Steiner stated that "the choice of Q_A is similar to defining the minimal clinically important effect in a clinical trial." For a fixed value of x_t, these mortality rates $p_0(x_t)$ and $p_A(x_t)$ are constants. By using the sequential probability ratio test (SPRT), the possible log-likelihood ratio score for patient t is:

$$W_t|x_t = \begin{cases} \log\left\{\dfrac{(1 - x_t + Q_0 x_t)Q_A}{(1 - x_t + Q_A x_t)Q_0}\right\}, & \text{if } y_t = 1, \\ \log\left\{\dfrac{1 - x_t + Q_0 x_t}{1 - x_t + Q_A x_t}\right\}, & \text{if } y_t = 0. \end{cases} \tag{3.11}$$

Suppose we set $Q_A > Q_0$ such that we are able to detect an increase in the mortality rate (deteriorated performance). It is known that the mortality risk for a patient has to be nonnegative. We note that for all positive values of the mortality risk for patient t, $W_t > 0$ if $y_t = 1$ and $W_t < 0$ if $y_t = 0$. Moreover, W_t is a decreasing function of x_t. This is illustrated in Figure 3.1. We can view

W_t as a penalty-reward score given to the clinical procedure, depending on the mortality risk and the outcome of mortality. If there is a mortality, the penalty will be large if the risk is small, and small if the risk is large. This is contrary to when there is no mortality. The reward given will be small if the risk is small, and large (negatively) if the risk is large.

This is similar when we set $Q_A < Q_0$ such that we are able to detect a decrease in the mortality rate (improved performance). If there is a mortality, the penalty will be large (negatively) if the risk is small, and small (negatively) if the risk is large. When there is no mortality, the reward given will be small (positively) if the risk is small, and large (positively) if the risk is large. As a decision rule, larger positive values of W_t tend to indicate stronger evidence against H_0 in support of H_A.

This illustrates that the log-likelihood ratio scores W_t for a patient proposed by Steiner et al. (2000) can be readily interpreted as a penalty-reward score given to a particular clinical procedure. As such, we will like to propose a set of test procedures which will alleviate the problem of interpretation of the test statistics, specifically if we have larger positive values of W_t for one clinical procedure as compared to another clinical procedure, it can be interpreted that the performances of the two clinical procedures are different. Moreover, this penalty-reward score is derived using the SPRT which is inspired from the classical likelihood ratio test. This likelihood ratio test is central to the famous Neyman and Pearson (1933) approach to statistical hypothesis testing.

As discussed earlier, the log-likelihood ratio score for patient t, W_t can be

viewed as a penalty-reward score given to each clinical procedure, depending on the mortality risk and the outcome of mortality. We will issue the penalty-reward score for patient t to Procedure i as:

$$W_{it}|x_{it} = \begin{cases} \log\left\{\dfrac{[1-x_{it}+Q_0(x_{it})\cdot x_{it}]Q_A(x_{it})}{[1-x_{it}+Q_A(x_{it})\cdot x_{it}]Q_0(x_{it})}\right\}, & \text{if } y_{it}=1, \\ \log\left\{\dfrac{1-x_{it}+Q_0(x_{it})\cdot x_{it}}{1-x_{it}+Q_A(x_{it})\cdot x_{it}}\right\}, & \text{if } y_{it}=0, \end{cases} \tag{3.12}$$

where $Q_0 = 1$ and $Q_A = \hat{Q}(x_{it})$. Upon the establishment of the penalty-reward system, there is a direct relationship between the penalty-reward scores W_{it} after accounting for the outcomes y_{it}, and the mortality risks x_{it} as $W_{it} = g_i(x_{it})$. From (3.12), by using Taylor series, it can be shown that $W_{it} = \beta_{i0} + \beta_{i1}x_{it} + \beta_{i2}x_{it}^2 + O(x_{it}^3)$. The use of the quadratic form for $g_i(x_{it})$ is further supported by the plots of penalty-reward scores against mortality risks in Figure 3.1. In order to compare the performance of two clinical procedures, one can consider comparing the following models:

$$W_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_{it}^2, \tag{3.13}$$

versus

$$W_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 x_{it}^2 + \beta_3 I(i=1) + \beta_4 x_{it} I(i=1) + \beta_5 x_{it}^2 I(i=1), \tag{3.14}$$

where $I(i=1)$ is an indicator function with a value 1 if we are appraising Procedure 1, or 0 if we are appraising Procedure 2. If there is no difference between the performances of the two procedures, (3.13) and (3.14) will yield the same model, with β_3, β_4 and β_5 all being equal to 0. But if their performances are indeed different, (3.13) and (3.14) will be different. In furtherance of the comparison,

one can utilize the earlier discussed deviance goodness-of-fit test (McCullagh and Nelder, 1989), where (3.14) is the full (or saturated) model and (3.13) is the reduced model. We note that $\hat{Q}(x_{it})$ is used in (3.12) as a known function, thus the calculated penalty-reward scores will still be conditionally independent on $\hat{Q}(x_{it})$ for both clinical procedures.

In the event that the type I error rate for the test is marginally different from the initial nominal level, it is possible that through some statistical adjustment to the significance level or p-value, we will be able to ensure that they will be compatible. For such technicalities, these are outside the scope of this paper. Similarly, addressing the concerns in the usage of the linear regression when it is performed on an unbalanced data set, Littell et al. (2002) stated the usefulness for unbalanced data as this method does not generally require balanced datasets.

SECTION 3. NUMERICAL EXAMPLES

To illustrate our proposed test procedures, as well as the proposed algorithm to estimate $p(x_t)$, an illustration of two examples using real data is shown. The proposed algorithm to estimate $p(x_t)$ is employed to visualize the form of the odds ratio and to give an illustration of the effectiveness of the clinical procedures or physicians' performance in each example. Our proposed test procedures and the McNemar's test are employed to detect any differences in performance in one of the examples and their corresponding results are presented.

Section 3.1 First Example: Acute Myocardial Infarction Admission in a Hospital

The first example utilizes data on the outcomes of patients with an acute myocardial infarction (more commonly known as heart attack) who are admitted to an anonymous hospital, collected as part of the NHS Research and Development funded EMMACE-1 (Evaluation of Methods and Management of Acute Coronary Events) Study (Dorsch et al. 2000). The post-operative outcomes after thirty days were collected for these patients admitted over a 3-month period. The mortality risk for each patient was both calculated and authenticated locally at the hospital. A total of 123 patients were observed and a cognizance of 27 deaths resulted in a mortality rate of 21.95%.

Since the mortality risk x_t is between 0 and 1 and from the previous studies of the mortality risk distribution, its theoretical model distribution may be modeled using a beta distribution. Quantile-quantile plots was then used to estimate the parameters among the probable location-scale family of beta distributions and the model distribution for the data set was parameterized by shape parameters $\alpha = 1$ and $\beta = 3$.

Supplementary to the discussion earlier, a scatter plot of the data $(x_t,\ y_t)$ will not be too informative. We are still able to note some characteristics of the data, namingly that the mortality rate appears to increase with the mortality risk and that the probable model distribution of the mortality risk, x_t is a decreasing probability distribution where smaller mortality risk values are more probable. To obtain more revelatory features of the data, we obtained a plot of the estimated

mortality rate $\hat{p}(x_t)$ against the mortality risk x_t, using kernel-based matching estimators in (3.4), in Figure 3.2. Subsequently, we can also transform the estimated mortality rate $\hat{p}(x_t)$ to map the odds ratio against the mortality risk x_t to obtain a more informative chart of the odds ratio function, as shown in Figure 3.2. From this plot in Figure 3.2, we observed that a horizontal line fit may be adequate as the points are randomly scattered around a horizontal line, showing no relationship between odds ratio and mortality risk, an indication that the effect of the true performance of the hospital might be the same (on the odds ratio) regardless of the mortality risk. In order to smoothen the estimates $\hat{p}(x_t)$, we apply the MSE criterion in (3.6) to find an estimate of the odds ratio Q and we obtain $\hat{Q} = 0.72$. Since $\hat{Q} < 1$, this suggests that there is possibly a decrease in the mortality rate across all levels of mortality risks. All in all, it supports the findings in Johnson et al. (2005) that the hospital exhibits "consistently good performance". Though Johnson et al. (2005) also found that there is a possible sudden downturn in the hospital's performance, our proposed method will consider all the available data as a whole, thus the deterioration in performance might be "averaged" (offset) by the other performances of the hospital. It is important for both clinicians and governance boards that our proposed method is not a monitoring tool but it is to obtain an overview of the average performance of the hospital accounting for the mortality risk x_t. It is more fundamental to reflect the overall quality of medical practice than to seek a possible occurrence of an isolated situation. We are not implying that such isolated situations should be neglected. They should still be identified for further investigations to reduce the variability in the quality of

medical practice and plausible causes for such situations.

Section 3.2 Second Example: Cardiac Surgery Operations in a Hospital

Our next example is exemplified with data from an anonymous hospital in UK. For this data set, the patients underwent cardiac surgery operations in the hospital and their post-operative outcomes after thirty days were collected. The given corresponding mortality risk x_t for each patient was both calculated and authenticated locally at the hospital. Due to confidentiality and anonymity of the data, only a subset of the data is in our illustration. A total of 426 patients over a period of time are considered and the data is stratified based on two physicians, with the first 322 patients being treated by a trainee physician, and the remaining 104 patients being treated by an experienced physician. The resulted unadjusted mortality rates of 2.80% and 3.85% are observed respectively, thus showing that the trainee physician is probably performing better than the experienced physician. However the standardized mortality ratio (SMR) for both surgeons are recorded as 0.952 and 0.683. This statistic is the ratio of observed mortality rate to predicted mortality rate. In our example, it indicates contradicting conclusions that the trainee physician is actually performing worse than the experienced physician. This suggests that adjustment for the patient mix is critical. In fact, from further examination of the data, it is found that the patient mix for both physicians are different, with the experienced physician treating patients of higher mortality risk.

Upon using the kernel-based matching estimators in (3.4), by pooling patients treated by both physicians, the plot of odds ratio against the mortality risk x_t in

Figure 3.3 seems to indicate that the odds ratio might not be the same across all levels of mortality risks, but instead it is more likely to be a linear function of the mortality risk x_t. As such, we apply the MSE criterion in (3.6) with $\hat{Q} = ax_t + b$ to find an estimate of the odds ratio Q function. We then obtain the plot of mortality rate $\hat{p}(x_t)$ against mortality risk x_t after smoothing for both physicians in Figure 3.3. Indeed it suggests that the trainee physician may be performing better than the experienced physician for patients of extremely low mortality risk (mortality risk below 0.0316) for the probable reason that the experienced physician might have taken over the operations for such patients if there are implications. This also account for a low number of such patients seen by the experienced physician. It can also be noted that the performance of the experienced physician is generally better than that of the trainee physician for patients across other levels of mortality risks (mortality risks between 0.0316 and 0.139). It is not meaningful to compare the performance of both physicians for patients of mortality risks above 0.139 as the trainee physician did not operate on such patients.

Subsequently, we conduct tests to evaluate the performance between the two physicians. The McNemar's test results in a p-value of approximately 1, upon the pairing of patients with the same mortality risk for both physicians. This additional step of pairing is not only troublesome, it is also cost-ineffective as we obtain only 94 pairs of patients as a result. This imply that 228 and 10 patients are left out in this analysis for the trainee physician and experienced physician respectively. There are also only 6 discordant pairs, which indicates that 88 concordant pairs are not utilized using the McNemar's test, thus possibly

losing valuable information from the data.

The implementation of the risk-adjusted test procedures proposed in the earlier Section yield both p-values of approximately 0 if we implemented the test procedures formulated using SPRT, based on the models discussed in (3.13) and (3.14), as well as that formulated using logistic regression, based on the logistic models discussed in (3.9) and (3.10). With our proposed method, we are also able to identify the extent of the differences in their performances for different values of the mortality risks, as shown in Figure 3.3 and it does suggest that the performance of the experienced physician is generally better than that of the trainee physician for patients across most levels of mortality risks (mortality risks between 0.0316 and 0.139).

It is important to note that the implementation of the proposed risk-adjusted test procedures are able to indicate probable differences in the performances of the physicians, controlling for all possible risk-adjusting factors mentioned above. Only upon seeking out these probable differences, there can begin a process to improve the performances, which may include retraining of staff or upgrading of equipment. For such technicalities, these are outside the scope of this paper.

SECTION 4. SIMULATION STUDY TO COMPARE MCNEMAR'S TEST AND PROPOSED METHODS

To further investigate the efficiency of our proposed test procedure, a simulation study was conducted. A comparison of our proposed procedure and the McNemar's test will be analyzed and their corresponding efficiencies will be presented. The focus of the comparison is on the following factors: size of the dif-

ference in parameters being tested for, the distribution of the mortality risk x_t and the odds ratio function assumed. The basis for determining the parameters and various aspects of the simulation study will be by simulating a data set with distributional characteristics which mimics that of the real data set in the first example, which we had discussed earlier in Section 3. This will ensure that our simulation studies are befitting of real-life scenarios.

Section 4.1 First Simulation Study: Under Constant Odds Ratio

The first part of our simulation study was to show the performance of our proposed procedure, as compared to the McNemar's test. The hypothesis of interest will be analogous to that for Steiner's risk-adjusted CUSUM, in which primary interest is with regards to the constant odds ratio of the mortality of a patient. To compare the performance of two clinical procedures, we will be testing this constant odds ratio, where $H_0 : Q_1 = Q_2$ versus $H_A : Q_1 \neq Q_2$, where Q_i is the odds ratio of the mortality of a patient under Procedure i ($i = 1, 2$). This is equivalent to testing $H_0 : \Delta Q = 0$ versus $H_A : \Delta Q \neq 0$ where $\Delta Q = Q_1 - Q_2$ is the difference in the odds ratio between the two clinical procedures. Based on the earlier data set, we will specify $Q_1 = 0.72$ as found in Section 3.

For each value of ΔQ considered in this study, the simulation was replicated 10000 times, resulting in 10000 data sets. Each data set comprised data for $n = 2000$ patients, with 1000 patients treated using Procedures 1 and 2 respectively. The mortality risk x_{it} for each patient was drawn from a beta distribution with shape parameters $\alpha = 1$ and $\beta = 3$, and the corresponding discrete outcome

generated from a Bernoulli distribution with $p_i = Q_i x_{it}/(1 - x_{it} + Q_i x_{it})$ where $Q_1 = 0.72$ and $Q_2 = 0.72 + \Delta Q$. The data will be analyzed to compare if the performance of the two clinical procedures are different. It is important to note that the McNemar's test will require an additional step before this can be implemented. In our study, the pairing of patients between Procedures 1 and 2 will be accomplished using a rule that will invoke the principle of optimal distance. This implies that the patients treated upon using each clinical procedure are matched based on their mortality risks and patients with the closest mortality risk for Procedures 1 and 2 are matched, with no patients (100% matching), 200 patients each (80% matching), 500 patients each (50% matching) and 800 patients each (20% matching) being left out in each matching procedure. This will adequately address a practical limitation of using the McNemar's test in which valuable information is possibly lost through the use of matched pairs, where some patients might not be taken into account. For example, in Chen, Connors and Garland (2008), due to a matching percentage of only 20%, information from the remaining 80% of the initially observed patients was not taken into account.

Our proposed test procedure will also be executed by using the test procedures formulated from SPRT on the models discussed in (3.13) and (3.14), as well as on the logistic models discussed in (3.7) and (3.8). In order to identify the importance of using the correct model of the relationship between $logit[p(x_t)]$ and x_t, we also performed the test on the logistic models discussed in (3.1) and (3.2).

Each procedure for testing will be carried out based on a test of the relevant null hypothesis at the 5% and 1% significance level, with the empirical power being

defined as the proportion of each 10 000 data sets in which the null hypothesis H_0 was rejected by the corresponding test.

First we examine the empirical type I error rates for the four tests in Table 3.1 and we see that our three tests and the McNemar's test have empirical type I error rates which are compatible with the nominal 5% and 1% (not shown here because similar conclusion is obtained) levels, with error rates for our tests minimally lower than the nominal levels and that for the McNemar's test (with matching percentage of 100%) marginally higher. We conclude that for the purpose of comparing two clinical procedures, all tests appear reasonable under the simulation settings considered here. As the matching percentage for the McNemar's test decreases, the empirical type I error rates are also observed to be decreasing.

Examining the empirical power of the tests, in broad terms, the performance between the McNemar's test (with matching percentage of 100%) and the test comparing the incorrect pair of models (3.1) and (3.2) is similar. The test procedures formulated from SPRT on the pair of models (3.13) and (3.14), as well as on the logistic models discussed in (3.7) and (3.8) also show similar performance. These findings are consistent when the tests are performed at both 5% and 1% significance level. Some of these findings are shown in Table 3.1. Though McNemar's test has performed adequately, it is important to note that the experimental setting required for this test is also more troublesome and less cost effective. As the McNemar's test requires patients from both clinical procedures to be matched, information is lost as there are patients not taken into account due to the matching procedure. This results in a loss in power of the McNemar's test, as shown

in Table 3.1. As the matching percentage decreases, the empirical power of the McNemar's test also decreases significantly.

Our proposed test procedures will not meet this drawback as it does not require that the patients to be matched, thus it is much more cost effective, with no compromise in the power of the test. Moreover, by looking at the empirical power of the tests comparing (3.7) and (3.8), and that comparing (3.1) and (3.2), we observe that the power of the test is diminished due to the incorrect specification of the logistic model, thus showing the importance of the introduction of the independent variable $logit(x_{it})$ in the logistic models.

We then repeat the above simulation study by changing the setting such that each data set comprised data for $n = 2000$, 1000, 500 and 200 patients, with 1000, 500, 250 and 100 patients treated using both Procedures 1 and 2 correspondingly. The data will be analyzed to compare if the performance of the two clinical procedures are different again.

First we examine the empirical type I error rates for the three tests in Table 3.2 and we see that our two tests and the McNemar's test have empirical type I error rates which are compatible with the nominal 5% level, with error rates for our tests minimally lower than the nominal levels and that for the McNemar's test (with matching percentage of 100%) marginally higher. We conclude that for the purpose of comparing two clinical procedures, all tests appear reasonable under the simulation settings considered here. We also notice that as the sample size for each data set decreases, the empirical type I error rates are also observed to be increasing.

Examining the empirical power of the tests, in broad terms, the performance between the McNemar's test and our two proposed tests is different, with the McNemar's test showing less power and both our tests showing similar but higher power as compared to the McNemar's test. These findings are shown in Table 3.2. Similarly, we also noticed that as the sample size for each data set decreases, the empirical power of the tests decreases significantly. This illustrate that all the tests are dependent on the sample size, similar to most test procedures.

Section 4.2 Second Simulation Study: Under Variable Degree in Performance Differences across Mortality Risks

In the second part of our simulation study, it is designed to investigate probable limitations in using the McNemar's test. The primary interest is the same as before, which is also to compare the performance of two clinical procedures, but the odds ratio is not taken to be identical (constant) across all levels of mortality risk x_t. Instead it is taken to be a linear function of the mortality risk x_t. The non-constant odds ratio is valid, as it is possible that the effect of the performance of the clinical procedure is different across the range of mortality risk x_t as shown by our second example, which we had discussed earlier in Section 3. The simulation process is similar to that conducted in the earlier experiment, but instead of constant Q we will specify $Q_1(x_t) = 0.30$ and $Q_2(x_t) = \alpha + \beta x_t$, listed in Table 3.3. Similarly, once the data set is obtained, it will be analyzed to compare the performance of the two clinical procedures by implementing the proposed test procedures, with the exception that instead of comparing (3.7) and (3.8), we will compare (3.9) and (3.10). The plots of the mortality rate $p(x_t)$ against the mortal-

ity risk x_t for the different odds ratio function $Q_2(x_t)$ considered in our simulation settings are also given in Figure 3.4.

Upon the examination of the empirical type I error rates for the four tests in Table 3.3, we observe that all tests have empirical type I error rates which are compatible with the nominal 5% level. The four tests appear to be appropriate for the comparison of two clinical procedures. However, through the examination of the empirical power of the tests, all the tests differ substantially in performance. Firstly if the effect between the performances of the two clinical procedures differs across the range of mortality risks, such as one procedure performing better for patients of lower mortality risks but worse for patients of higher mortality risks, the power of the McNemar's test will be greatly undermined. This is particularly so when the clinical procedures' plot of mortality rate, $p(x_t)$ against the mortality risk x_t intersect, as shown in Figure 3.4. The power of the McNemar's test can be significantly lower than that of our proposed test procedures. The primary reason is due to the cancellation of the effects between the performances of the two clinical procedures, since one procedure performs better than the other across a range of mortality risks but tend to perform worse over the other range of mortality risks. This scenario is befitting of real-life situations. For example in the treatment of coronary heart disease, coronary angioplasty is recommended for patients of lower risk and might not be as effective for patients of higher risk, while CABG surgery is conducted on patients of higher risk and might not be as effective for patients of lower risk. By looking at the empirical power of the tests comparing (3.9) and (3.10), and that comparing (3.1) and (3.2), we again observe that the power of the

test is diminished due to the incorrect specification of logistic model, thus showing the importance of the introduction of the independent variable $logit(x_{it})$ in the logistic models.

In summary, on top of the drawbacks of high costs of implementation, the McNemar's test is not able to identify the point of intersection in the clinical procedures' plot of mortality rate against mortality risk, since upon pairing of the patients, information of the patients' mortality risks are "lost". Using the proposed algorithm to estimate $p(x_t)$, we are able to obtain a more informative plot of mortality rate $\hat{p}(x_t)$ against mortality risk x_t after smoothing for each clinical procedure, thus identifying the probable point of intersection or the change in the effects between the performances of the clinical procedures. The proposed test procedures can also be implemented to test for the comparison of the two clinical procedures, while accounting for possible varying degree of the differences between the performances of the two clinical procedures across the range of mortality risks.

Section 4.3 Third Simulation Study: Under changes in the Underlying Mortality Risk Distribution

Rogers et al. (2004) have voiced their concerns about the effect of changes in the underlying mortality risk distribution on the performance of the risk-adjusted CUSUM chart. In our first example, the mortality risk distribution was modeled using a beta(1,3) distribution but this is not always the case. For clinical procedures, such as CABG surgery, they are performed on patients of higher risk. This will in turn result in a mortality risk distribution that is more skewed to the right, with more patients with higher risks. For other procedures, such as coro-

nary angioplasty which is recommended for patients of lower risk, the mortality risk distribution will be less skewed to the right, with more patients with lower risks. As such, this part of our simulation study is designed to investigate the sensitivity of our proposed test procedures to changes in the underlying mortality risk distribution for the two clinical procedures. Here the simulation process is similar to that conducted in the first experiment but the mortality risk for each patient for both procedures was drawn from a beta distribution with shape parameters $\alpha = 1$ but with different values of β. Once the data set is obtained, it will be analyzed to compare the performance of Procedures 1 and 2 by implementing the test procedures on models discussed in (3.13) and (3.14), as well as on the logistic models discussed in (3.7) and (3.8).

Similarly, we examine the empirical type I error rates for our proposed test procedures in Table 3.4 at nominal 5% level, we observe that the empirical type I error rates decrease as the distribution becomes less skewed to the right. This is due to the reason that as the distribution becomes less skewed to the right, the mortality risks of the patients become more concentrated at the lower risk values. As a result, the differences between the two procedures will be estimated more accurately around these risk values, thus leading to less evidence of any differences between the two procedures.

We conclude that as the distribution becomes less skewed to the right (when β increases), the tests become more conservative (that is, less likely to reject H_0 whether true or false). In contrast, as the distribution becomes more skewed to the right (such as when β decreases), the tests become more liberal (that is, more

likely to reject H_0 whether true or false).

In summary, the degree of conservativeness of the test, and therefore its validity and the conclusions drawn from the tests have to be treated with caution as well. It is possible that through some statistical adjustment to the significance level, we will be able to ensure that the type I error rates will be compatible to the initial nominal level.

SECTION 5. CONCLUSION

At present, risk-adjusted analytical tools are best used as a screening or monitoring procedure (Cook et al. 2008), rather than to compare the performances of clinical procedures. To compare performances of clinical procedures, it is common to obtain two groups of patients who were treated by either clinical procedures and their pre-operative mortality risks are matched. This ensures that the matched pairs are well-balanced with respect to all potentially confounding variables. Afterwhich, the McNemar's test is used to assess the difference between the two procedures. The McNemar's test has the advantage of being robust against the underlying mortality risk distribution as it has accounted for the risk distribution through the matching procedure. Due to the earlier mentioned, the McNemar's test will also not be able to show the exact performance of each clinical procedure as it is only able to establish differences between the two correlated proportions. It is also noted that the implementation of the McNemar's test becomes more troublesome and less cost effective as this might result in excessive loss in information because there might be many patients that are not paired and thus not taken

into account. This results in possibly significant loss in power of the McNemar's test, as shown in our simulation study. The McNemar's test also focuses only on the discordant pairs, thus possibly losing further valuable information from the data. Through our simulation study, we also demonstrate that the McNemar's test is highly sensitive to the degree of the differences between the performances of the two clinical procedures, for example if one is performing better for patients of lower mortality risks but worse for patients of higher mortality risks. This is because the McNemar's test assumes that the degree of the differences between the performances of the two clinical procedures are the same for any pair of patients.

It is noted from current literature, that in order to compare performances of clinical procedures, it is common to compare logistic models which utilize a linear function of the independent variable (mortality risk), using logistic regression. As such, this method is dependent on the model used. If a wrong model is used, this will in turn result in a less powerful test, as shown by our simulation studies. However, unlike linear regression, there is no R^2 associated with a logistic model, thus it is not simple to evaluate whether a model is wrongly used. It is then of importance that we develop a model-free technique to estimate the actual mortality rates for all levels of predicted mortality risk. This proposed diagnostic technique will allow us to evaluate the effectiveness of the clinical procedure by investigating the mortality rates and resulting odds ratio function against the mortality risks. More importantly, it does not assume any relationship between the mortality rates and risks.

Using the estimated mortality rates obtained utilizing our diagnostic tech-

nique, as well as using logistic regression, we present two sets of risk-adjusted test procedures that can also be used to compare the clinical procedures' performances, or specifically compare if the odds ratio function of the mortality risk is the same for both procedures. For the first set, we utilize the logistic regression to formulate the procedures by introducing the independent variable $logit(x_{it})$ in the logistic models, while for the second set, we proposed the risk-adjusted penalty-reward scores to reflect the pre-operative mortality risk and mortality outcome of each patient. This alleviates the problem of interpretation since larger positive values of W_t for one clinical procedure as compared to another clinical procedure can be interpreted that the performances of the two clinical procedures are different. It also provides an alternative test which is derived using the SPRT and inspired from the classical likelihood ratio test. This approach provides an intuitive way to average "evidence" over the patients, while adjusting for the case-mix of patient characteristics that might significantly affect the mortality risk. Both proposed test procedures also do not require that the patients to be matched, thus it is much more cost effective.

Our proposed test procedures are also able to overcome the drawback of being sensitive to the degree of the differences between the performances of the two clinical procedures, and are yet able to achieve good efficiency as compared to when the McNemar's test is used. It is important to note that similar to most tests, the degree of conservativeness of our proposed tests, and therefore its validity and the conclusions drawn from the test procedures have to be treated with caution. Finally, the last key advantage of our proposed test procedures is that

through slight modifications, it can also be used to compare more than two clinical performances.

Table 3.1. Empirical type I error and power at a 5% significance level under $H_0 : Q_1(x_t) = Q_2(x_t)$ versus $H_1 : Q_1(x_t) \neq Q_2(x_t)$, with the distribution of the mortality risk as beta(1,3)

n	$Q_1(x_t)$	$Q_2(x_t)$	McNemar's test with matching percentage of				Logistic Regression		SPRT
							Usual Models	Correct Models	
			100%	80%	50%	20%	(3.1) vs (3.2)	(3.7) vs (3.8)	(3.13) vs (3.14)
1000	0.72	0.72	0.0591	0.0419	0.0404	0.0328	0.0453	0.0456	0.0454
	0.72	0.30	0.9996	0.9990	0.9809	0.6164	1.0000	1.0000	1.0000
	0.72	1.12	0.9399	0.9038	0.6757	0.2783	0.9330	0.9632	0.9621

Table 3.2. Empirical type I error and power at a 5% significance level under $H_0 : Q_1(x_t) = Q_2(x_t)$ versus $H_1 : Q_1(x_t) \neq Q_2(x_t)$, with the distribution of the mortality risk as beta(1,3) for various n

n	$Q_1(x_t)$	$Q_2(x_t)$	McNemar's test (100% matching)	Logistic Regression Correct Models [(3.7) vs (3.8)]	SPRT (3.13) vs (3.14)
1000	0.72	0.72	0.0591	0.0456	0.0454
	0.72	0.3	0.9996	1.0000	1.0000
	0.72	1.12	0.9399	0.9632	0.9621
500	0.72	0.72	0.0613	0.0478	0.0472
	0.72	0.3	0.9648	0.9742	0.9703
	0.72	1.12	0.8216	0.8354	0.8338
250	0.72	0.72	0.0641	0.0491	0.0488
	0.72	0.3	0.8567	0.8930	0.8854
	0.72	1.12	0.4298	0.4734	0.4701
100	0.72	0.72	0.0679	0.0519	0.0517
	0.72	0.3	0.4931	0.5300	0.5269
	0.72	1.12	0.2158	0.2321	0.2302

Table 3.3. Empirical type I error and power at a 5% significance level under $H_0 : Q_1(x_t) = Q_2(x_t)$ versus $H_1 : Q_1(x_t) \neq Q_2(x_t)$, with the distribution of the mortality risk as beta(1,3) for true non-constant Q_2

$Q_1(x_t)$	$Q_2(x_t)$	McNemar's test	Logistic Regression		SPRT
			Usual Models	Correct Models	
			(3.1) vs (3.2)	(3.9) vs (3.10)	(3.13) vs (3.14)
0.3	0.3	0.0535	0.0446	0.0496	0.0495
0.3	$0.2 + 0.5\ x_t$	0.0708	0.3599	0.4960	0.4543
0.3	$0.8\ x_t$	0.5097	0.6431	0.7934	0.7638

Table 3.4. Empirical type I error rates of the test procedures for $\Delta Q = 0$ corresponding to various underlying mortality risk distributions for both clinical procedures under $H_0 : Q_1(x_t) = Q_2(x_t)$ versus $H_1 : Q_1(x_t) \neq Q_2(x_t)$

Distribution	Logistic Regression	SPRT
	(3.7) vs (3.8)	(3.13) vs (3.14)
Beta(1, 2.0)	0.0521	0.0478
Beta(1, 2.5)	0.0491	0.0467
Beta(1, 3.0)	0.0456	0.0454
Beta(1, 4.0)	0.0449	0.0440
Beta(1, 5.0)	0.0437	0.0439
Beta(1, 10.0)	0.0413	0.0412
Beta(1, 20.0)	0.0392	0.0395
Beta(1, 50.0)	0.0370	0.0384

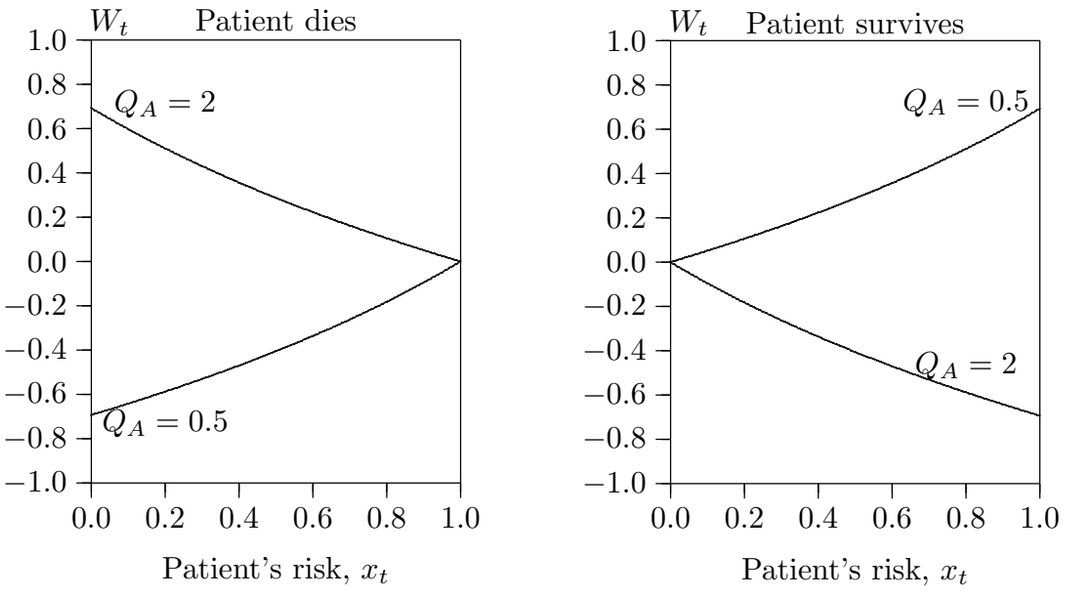

Figure 3.1. Penalty-reward score W_t awarded to a surgeon according to a patient's pre-operative risk x_t, where $H_0 : p_0(x_t)/[1 - p_0(x_t)] = Q_0 x_t/(1 - x_t)$ versus $H_A : p_A(x_t)/[1 - p_A(x_t)] = Q_A x_t/(1 - x_t)$.

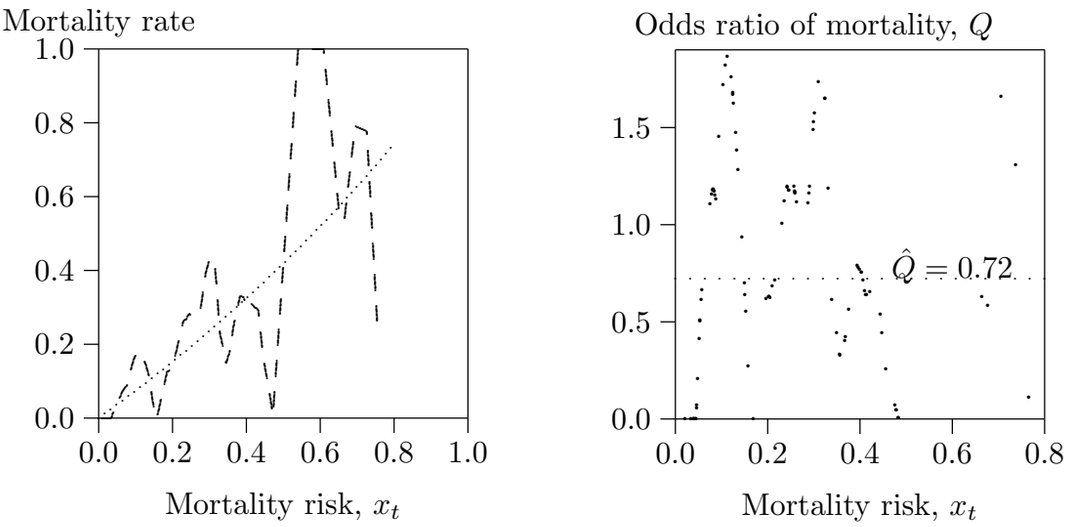

Figure 3.2. Plot of mortality rate $\hat{p}(x_t)$ against mortality risk x_t, and plot of odds ratio of mortality Q against mortality risk x_t after smoothing with patients with an acute myocardial infarction who are admitted to an anonymous hospital, collected as part of the EMMACE-1 Study. The dashed and dotted curves in the plot on the left represents the mortality rate before smoothing and after smoothing respectively.

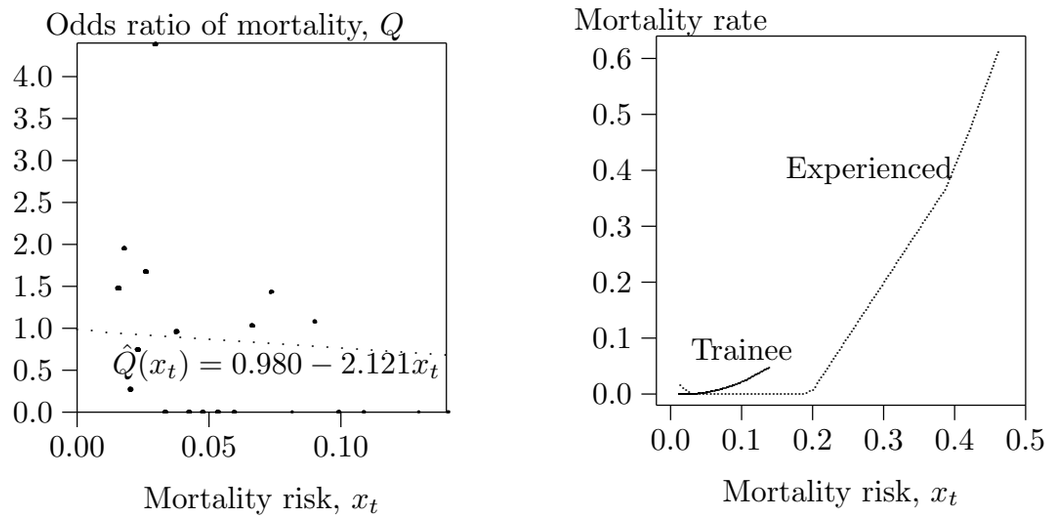

Figure 3.3. Plot of odds ratio of mortality Q against mortality risk x_t after smoothing and plot of mortality rate $\hat{p}(x_t)$ against mortality risk x_t, for trainee physician and experienced physician after smoothing for patients who underwent cardiac surgeries in an anonymous hospital in UK.

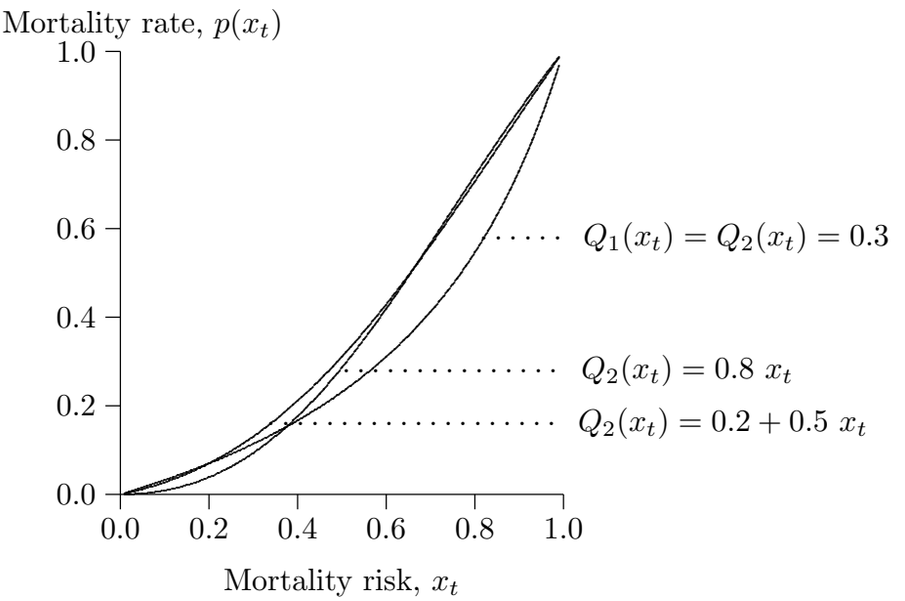

Figure 3.4. Plot of mortality rate $p(x_t)$ against mortality risk x_t.

CHAPTER 4: STANDARDIZED MORTALITY RATIO (SMR): FACTS AND MYTHS. A REVIEW ON THE USAGE OF SMRs

SUMMARY

The ability to assess medical practice is central to quality assurance. One such widely-used overall quality indicator and measurement tool will be the standardized mortality ratio (SMR) by comparing the observed mortality rate to the predicted rate. Despite being available for some time, health service providers are still skeptical on its ability to truly identify poor-quality providers. Recent paper has emphasized the validity of case mix adjustment methods used to predict the mortality rates. Beyond this methodological bias, in this paper, we will investigate various limitations of using the SMR by using worked examples and simulation studies. We also provide theoretical estimates of the mean and variance for SMR. The highlights of various possibly wrong interpretations through the use of SMR are also adequately discussed.

SECTION 1. INTRODUCTION

The evolution of the assessment of medical practice has been speeding up tremendously, as seen from recent literature (Werner and Bradlow, 2006, Clarke and Oakley, 2007, Krumholz et al., 2008, Biswas and Kalbfleisch, 2008, Steiner and Jones, 2009). Measuring quality of care in medical practice is a key component in improving efficiency in health care. The crux of monitoring clinical performances can be attributed to distinguished names from earlier times, such as Florence Nightingale, Ernest Codman and Lord Moynihan (Chambler and Emery, 1997, Kaska and Weinstein, 1998, Spiegelhalter 1999). Such assessment is playing an increasingly prominent role in quality management. For example, in 1999, an independent body, the UK National Institute of Clinical Excellence was established, after the UK General Medical Council found three doctors possibly guilty of professional misconduct over the quality of their heart surgeries conducted. The professional misconduct led to 29 mortalities out of 53 children who were operated at the Bristol Royal Infirmary (2001, BBC News 1998). In 2006, a pro tempore closure of the cardiac surgical department at the Radboud University in the Netherlands was also initiated after an analysis of the mortality rates was conducted. The reactions to these health care tragedies stress the importance of clinical performance monitoring as timely signals of deteriorated performance can be used to identify plausible causes and this will in turn avoid future avertible mortalities or other adverse health issues.

In order to make clinical performance information readily accessible to the public, payers and providers of care, the United States Centers for Medicare and

Medicaid Services has been collaborating with various health care organizations to participate in the Hospital Quality Alliance (2006). It is therefore crucial that information released is reasonably accurate and fairly representative such that it is of significant value. But the release of such performance report cards might lead to misinterpretation of the data. Patients in hospitals tend to differ notably in terms of pre-procedural risk of failure, which in this paper we will refer to as mortality risk. If this variability in the mortality risk is not taken into account when assessing a particular hospital's performance, this variability might result in additional fluctuation in the outcomes, thus masking the effect of the true performance of the hospital, and resulting in misapprehension of the true situation. For example, if a particular hospital has a relatively low mortality rate, it will give an impression that this hospital is highly, and vice versa. As such, the New York State Department of Health (2008) do not just publish raw mortality rates, they also publish risk-adjusted mortality rates, which is an indication of what a hospital's mortality rate would have been, had she treated patients identical to the state's average.

Another one such widely-used overall quality indicator and measurement tool will be the standardized mortality ratio (SMR). This statistic is calculated as:

$$\begin{aligned} SMR &= \frac{\text{Observed mortality rate}}{\text{Predicted mortality rate}} \\ &= \frac{\text{Number of observed deaths}}{\text{Predicted number of deaths}}, \end{aligned} \tag{4.1}$$

where the predicted mortality rate is calculated using a reference model, and this statistic serves as an indirect method of adjusting for the risk. For a hospital, a SMR value that is equal to 1 suggests that there is no difference between the

observed mortality rate and the predicted mortality rate. However, a SMR value greater or lesser to 1 suggests that the hospital's mortality rate is higher or lower than the predicted mortality rate respectively, thus also indicating whether the performance is good or poor respectively. This statistic has been available and been publicly released for some time and it is usually recommended to be viewed in context with other quality indicators. The SMR is used only as a tool to help health service providers identify the trends in their hospital mortality rate and make quality improvements based on the results. This practice has been implemented in various countries such as the United States, the United Kingdom, Sweden and Holland. For example, by tracking SMR and implementing a range of improvements as a result of what is learnt, the Walsall hospital in the United Kingdom was able to reduce mortality by 40% in only 4 years. Despite its usefulness, health service providers are still skeptical on its ability to truly identify poor-quality providers and they doubt its validity to be used as a measure for comparing providers publicly (Consortium of Chief Quality Officers, 2009).

In the literature to compare health care providers, there has always been concerns on the non-standardized documentation and coding of patients' conditions, as well as on the validity of case mix adjustment methods used to predict mortality rates. Beyond this methodological bias, on the other hand, Rogers et al. (2004) expressed their concerns about the effect of changes in the underlying mortality risk distribution on the performance of the risk-adjusted CUSUM chart. We are motivated to investigate the effectiveness of using SMR to compare hospitals when their corresponding mortality risk distributions are different. We demonstrate this

using a real data set. For this data set, the patients underwent cardiac surgery operations in an anonymous hospital in UK and their post-operative outcomes after thirty days were collected. The corresponding mortality risk for each patient was both calculated and authenticated locally at the hospital. The data is stratified based on 3 physicians, which we will identify them as Physicians A, B and C. We calculate the SMR values for Physicians A, B and C to be 0.6780, 0.8112 and 0.9318 respectively. This leads to a suspectible conclusion that the physicians differ in their performances, with Physician A showing the best performance and Physician C showing the worst performance amongst the three. Could this conjecture be flawed due to other reasons? Investigating the average mortality risk of patients seen by each physician, that for Physicians A, B and C are found to be 0.0309, 0.0439 and 0.0478, with Physician A operating on more patients of lower risks and Physician C operating on more patients of higher risk. A physician who operates on patients of lower risks should show better performance than another who operates on patients of higher risks by having a lower mortality rate. As such, the discrepancies in SMR values are possibly due to differences in the mortality risk distributions. In fact, using an alternative statistics in (4.6) proposed in Section 3, the performances between the 3 physicians are found to be similar with $\tilde{Q} = 0.5984$, 0.6008 and 0.6014.

In order to better understand the validity of using SMR to compare hospitals, in Section 2, we will provide a review of literature discussing various limitations of using SMR, as well as highlight various possibly wrong interpretations through the use of SMR. Theoretical estimates of the mean and variance of SMR are also

provided. An alternative statistic for the comparison of hospitals will be proposed in Section 3. Using real applications in health care context, the comparison of using the proposed method and SMR will be analyzed and the findings presented in Section 4. The conclusions and important findings will then be highlighted in the last section.

SECTION 2. FACTS AND MYTHS OF SMR

Section 2.1 Literature Review of Using SMR

There have been controversies in the inferences drawn from a wide variation in hospitals' SMR published publicly. However, these published SMR portray diversities in quality of care. In the 2007 hospital guide for UK, Dr Foster Intelligence depicted SMR as "an effective way to measure and compare clinical performance, safety and quality." But it has been noted that SMR requires the predicted mortality rate to be calculated and this is done using some reference model obtained from some developmental data set. Jones, Redmond and Templeton (1995) suggested that the mortality prediction models can be applied to a new data set only if the mortality risk distribution for the new data set is not statistically different from that of the developmental data set. We note that since the predicted mortality rate in SMR is calculated using some reference model, we emphasize that SMR is a comparison of the index data set with the developmental data set. As such, if the data tend to differ between each health care provider, the comparison tends to be undermined, especially if comparisons are done across large numbers of providers since it is unlikely that the data is similar.

Lilford et al. (2004), Iezzoni (1997) and Mohammed et al. (2009) assessed the validity of various case mix adjustment methods used to obtain the predicted mortality rate and identified the importance of using the correct underlying case mix adjustment method before any inferences on the comparison of health care providers can be made. Much literature focus on limitations of current case mix adjustment methods (Moore et al. 2010) due to inadequate risk adjustment and non-standardization of documentation and coding in administrative data. In the commentary released (Consortium of Chief Quality Officers, 2009), it was highlighted that when health care providers review the medical records of patients among those whose mortality is "higher than expected", the findings usually include finding very sick patients whose "expected" death rate is under-estimated. However, amidst the above discussed disadvantages of using SMR, there are still evident advantages in its usage, such as age-specific numbers of deaths are not required in its calculation and its robustness to violations of the assumption of proportionality. As such, Jarman et al. (2010) still promotes SMR as a powerful tool to assess quality of care in Holland.

In the literature to compare health care providers, there has always been concerns on the non-standardized documentation and coding of patients' conditions, as well as on the validity of case mix adjustment methods used to predict mortality rates. Beyond the above discussed methodological bias, suppose that the concerns on the non-standardized documentation and coding of patients' conditions are addressed, as well as the case mix adjustment methods used to predict mortality rates are correct, we are motivated to investigate the effectiveness of

using SMR to compare hospitals. Does that indicate that the use of SMR should be encouraged amidst the absence of this methodological bias? We will first derive theoretical estimates of the mean and variance of SMR.

Section 2.2 Theoretical Estimates of Mean and Variance of SMR with Discussion on Limitations of Using SMR

Suppose we let y to be the post-procedural outcome for a patient, it corresponds to one of two possible outcomes (success or failure). We assume y_t is the outcome for patient t. (1 if there is a mortality or 0 if a patient survives after implementation of the clinical procedure). Notice that the outcome of the clinical procedure may not be observed immediately after its implementation, with one example being that for cardiac operations in which the outcome of mortality is usually determined within 30 days from surgery. If patient t dies anytime within 30 days from the surgery, y_t will be assigned a value of 1 and if the patient survives after 30 days from surgery, y_t will be assigned a value of 0. As a result, we have the following probability function of y_t, $f(y_t|p) = p^{y_t}[1-p]^{1-y_t}$, where p is the mortality rate.

We further assume x_t to be the mortality risk for patient t and it is estimated prior to the implementation of the clinical procedure and it depends on the risk factors present for the patient. This risk can be determined by using a rating method, such as Parsonnet risk factors (Parsonnet, Dean and Bernstein, 1989) for cardiac operations. Afterwhich, a logistic regression model is used to convert these scores obtained from the rating method, to a risk value between 0 and 1. The risk may also be computed based on a logistic regression model fitted to sample data

or past data set, such as the EuroSCORE (Nashef et al., 1999) which is used to evaluate the risk of patients for cardiac operations.

Since the mortality risk x_t is between 0 and 1, and from previous studies of the mortality risk distribution, its theoretical model distribution may be modeled as beta(α,β). Morgan and Henrion (1990) and Moitra (1990) pointed out that the use of beta distributions in modeling data in the form of proportions is due to its large variety of shapes. Moreover, Hakes and Viscusi (1997) also discussed that since the beta distribution is flexible and can assume a vast variety of skewed and symmetric shapes, its use is not notably restrictive. As such, by assuming that the risk distribution is modeled as beta(α,β), the theoretical estimates of the mean $E(SMR)$ and variance $Var(SMR)$ of SMR are derived as:

$$\begin{aligned} E(SMR) &= E\left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}\right) \\ &= \left(\frac{Q\alpha}{\beta+Q\alpha} + \frac{Q(1-Q)\alpha\beta(\alpha+\beta)}{(\alpha+\beta+1)(\beta+Q\alpha)^3}\right)\left(\frac{\alpha+\beta}{\alpha}\right) - \left(\frac{Q\beta(\alpha+\beta)^2}{\alpha(\alpha+\beta+1)(\beta+Q\alpha)^2 n}\right) \\ &+ \left(\frac{Q\alpha}{\beta+Q\alpha} + \frac{Q(1-Q)\alpha\beta(\alpha+\beta)}{(\alpha+\beta+1)(\beta+Q\alpha)^3}\right)\left(\frac{\alpha+\beta}{\alpha}\right)^3\left(\frac{\alpha\beta}{n(\alpha+\beta+1)(\alpha+\beta)^2}\right), \end{aligned} \tag{4.2}$$

$$\begin{aligned} Var(SMR) &= Var\left(\frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}\right) \\ &= \left(\frac{Q)(\alpha+\beta)}{\beta+Q\alpha} + \frac{Q(1-Q)\beta(\alpha+\beta)^2}{(\alpha+\beta+1)(\beta+Q\alpha)^3}\right)^2 \frac{1}{n}\left(\frac{\beta}{\alpha(\alpha+\beta+1)}\right. \\ &+ \left(1 - \frac{Q\alpha}{\beta+Q\alpha} + \frac{Q(1-Q)\alpha\beta(\alpha+\beta)}{(\alpha+\beta+1)(\beta+Q\alpha)^3}\right)\left(\frac{Q\alpha}{\beta+Q\alpha} + \frac{Q(1-Q)\alpha\beta(\alpha+\beta)}{(\alpha+\beta+1)(\beta+Q\alpha)^3}\right)^{-1} \\ &\left. - \frac{2Q\beta(\alpha+\beta)}{(\alpha+\beta+1)(\beta+Q\alpha)^2}\left(\frac{Q\alpha}{\beta+Q\alpha} + \frac{Q(1-Q)\alpha\beta(\alpha+\beta)}{(\alpha+\beta+1)(\beta+Q\alpha)^3}\right)^{-1}\right), \end{aligned}$$

(4.3)

where Q is the odds ratio of the mortality of a patient and n is the sample size. If $Q = 1$, this indicates that the estimated risk x_t is the same as the mortality rate $p(x_t)$ for all x_t. We note that x_t is based on the current conditions before taking into account the effect of the true performance of the hospital. As such, there is no difference in the performance of the hospital before and after taking into account the effect of the true performance of the hospital. As Q increases, the mortality rate increases, thus showing that there is a deterioration in performance. If one hospital has a smaller Q than another hospital, it indicates that the former hospital is performing better and vice versa.

The theoretical estimates of the mean and variance of SMR is compared with that obtained from simulation studies and they are found to be compatible (not shown here because similar conclusions are obtained). From the theoretical estimate of $Var(SMR)$, we also observe that as $n \to \infty$, $Var(SMR) \to 0$. Cook et al. (2008) also highlighted that a confidence interval should always be included to illustrate the precision of the SMR estimate. Upon obtaining theoretical estimates of the mean and variance of SMR, we can obtain a confidence interval for the SMR estimate. For such technicalities, these are outside the scope of this paper.

Using the theoretical estimates of the mean for SMR, we will discuss beyond the earlier discussed methodological bias, the effectiveness and limitations of using SMR to compare hospitals when their corresponding mortality risk distributions are different. Though it has been discussed that SMR should not be used to compare institutions with different patient case mix (Julious, Nicholl and

George, 2001, Breslow and Day, 1987, Glance, Osler and Shinozaki 2000), these discussions were based on simulation studies and real examples. In Glance, Osler and Shinozaki (2000), they presented that SMR will decrease in a linear fashion as the average mortality risk increases. However, we will like to highlight out some important points in their simulation study. Firstly, from their analysis, SMR should decrease in a quadratic fashion as the average mortality risk increases, as the quadratic model will offer a better fit. Secondly, this relationship of SMR decreasing with the increase in the average mortality risk is not necessarily true. It depends on how the patient case mix is selected from their original data set, in which across the 10 deciles of risk, the odds ratios, or more generally the performance of the institution is not the same.

In this paper, we will present a more complete and theoretical approach to this investigation.

Firstly, past literature has shown that the average mortality risk of patients undergoing cardiac surgery has been predominantly increasing over the years. Parsonnet, Bernstein and Gera (1996) demonstrated that when the Parsonnet model was applied to the patients of the Beth Israel Hospital in Newark, the average mortality risk of patients had progressively increased by 47.7% in 1994, as compared to that of 6.5 in 1988. The National Adult Cardiac Surgical Database Report (2001) also showed similar trends when the average mortality risk of patients in 1999 had increased by 20% over 3 years. Since much literature has placed great emphasis on the average mortality risk, we will conduct the discussion, referring differences in the risk distribution to differences in the average mortality risk.

The risk distribution will be modeled by the beta distribution with shape parameters $\alpha = 1$ but with different values β, resulting in average mortality risks of $\alpha/(\alpha+\beta) = 1/(1+\beta)$, with $n = 1000$. We will also consider various values of the odds ratio where $Q = 1/2$, $2/3$, 1, 1.5 and 2, which are some values of proposed odds ratio of interest (Steiner et al. 2000, Novick et al. 2006, Matheny, Ohno-Machado and Resnic 2007). The plot of $E(SMR)$ against the average mortality risk for these values of odds ratio is shown in Figure 4.1.

Firstly, we observe that if the odds ratio $Q = 1$, regardless of whether the mortality risk distributions are different, $E(SMR)$ is always equal to 1. It is robust to differences in the risk distribution.

However, for $Q = 1/2$, $2/3$, 1.5 and 2, $E(SMR)$ is affected by differences in the risk distribution. Specifically, if $Q > 1$, $E(SMR)$ decreases towards 1 as the average risk increases, and $E(SMR)$ is always above 1, but if $Q < 1$, $E(SMR)$ increases towards 1 as the average risk increases, and $E(SMR)$ is always below 1. This indicates that if one hospital has a SMR value greater than 1 while another has that smaller than 1, we can compare these two hospitals and conclude that the latter is performing better than the former.

However, if both hospitals have SMR values greater than 1 or smaller than 1, we are not able to compare these two hospitals without looking at the risk distribution. This can be seen from Figure 4.1. For example, we see that the $E(SMR)$ values for $Q = 2$ at higher average risks, are smaller than that for $Q = 1.5$ at smaller average risks. This might lead us to conclude that the former is performing better than the latter based on the $E(SMR)$ values. But we note that

for the former, $Q = 2$ which is greater. As such, the earlier conclusion obtained will be erroneous. This is also similar when $Q < 1$ in which the $E(SMR)$ values for $Q = 1/2$ at higher average risks, are higher than that for $Q = 2/3$ at smaller average risks. This might again lead us to conclude that the former is performing worse than the latter based on the $E(SMR)$ values. But we note that for the former, $Q = 1/2$ which is smaller, thus it is supposed to be performing better. As such, a contradiction is reached with the earlier conclusion to be incorrect.

It has discussed that if the average mortality risk increases, SMR values are expected to drop due to the decrease in the predicted number of deaths. However, we will like to highlight that the observed number of deaths will also be in turn affected. In conclusion, the E(SMR) values do depend on both the true underlying performances of the hospitals and the mortality risk distribution of the patients seen at each hospital, as seen from our theoretical study here.

Section 2.3 Possibly Wrong Interpretations through the use of SMR

Upon identifying various limitations of using SMR, as well as that SMR depend on both the true underlying performance and the case mix of patients, we will highlight some possibly wrong and proper interpretations through the use of SMR in the literature.

Amin et al. (2006) studied the roots of mortality after diagnosis of hepatitis B or hepatitis C infections through the use of a large community-based study. It was subsequently found that for liver-related deaths, the SMR for hepatitis B, hepatitis C, and hepatitis B and C co-infected patients were 12.2, 16.8 and 32.9.

The conclusion that all 3 groups of patients had increased risk of liver-related death compared with the standard population is correct. However, it was also highlighted that the greatest excess occurs in people diagnosed with hepatitis B and C co-infection. We note that despite having a highest SMR value for this group of patients, information on the mortality risk distribution is not released. As discussed earlier, if the average mortality risk for this group of patients is the lowest, it is possible that this co-infection might not be "performing that badly", thus indicating that people diagnosed with co-infection might not have the highest odds ratio of mortality.

Huber-Wagner et al. (2009) studied the effect of whole-body CT during trauma resuscitation on survival through the use of a retrospective, multicentre study. Because of the use of a multicentre study, the data might differ between each centre, thus the case mix adjustment model might not valid.

Looking at the findings, SMR based on the trauma and injury severity score (TRISS) was 0.745 for patients given the whole-body CT while for those given a non-whole-body CT, SMR was 1.023. Similar results were also obtained using revised injury severity classification (RISC) score with the SMR for the former to be 0.865 versus 1.034 for the latter. The conclusion was that the integration of whole-body CT significantly increased the probability of survival in patients with polytrauma. This conclusion is correct because we are comparing SMR, with one below 1 and the other above 1. This is because in the earlier section, we conclude that regardless of the average risk, if $Q > 1$, $E(SMR)$ is always above 1, and if $Q < 1$, $E(SMR)$ is always below 1. This conclusion will not be valid if both SMR

are above or below 1, even if one is higher than the other.

In 2009, there was a study to investigate the relationship between mortalities and prognostic factors in anorexia nervosa (Papadopoulos et al 2009). Some conclusions are highlighted. Firstly, it was mentioned that the highest SMR of 650.0 was found among patients with anorexia nervosa, with the second highest SMR of 18.9 was for psychoactive substance use. The direct comparison of the value of SMR should not done especially since the mortality risk distribution in each group of patients is not discussed. Moreover, it was further discussed that a very high SMR was found for the first year after first hospitalisation for anorexia and the SMR was still significantly high 20 years or more after the first hospitalisation. This notion of being significantly high should be highlighted that the basis of comparison is with the general population.

Lastly, there was a discussion on the reduction in acute myocardial infarction mortality in the United States from 1995 to 2006 (Krumholz et al. 2009). The same case mix adjustment model was used across all the years. Past literature has shown that the average mortality risk of patients undergoing cardiac surgery has been predominantly increasing over the years. As such, the model might have changed over the years. The mortality risk distribution will also have been different with different profile of patients over the years, thus the calculated SMR should not be compared.

SECTION 3. GENERAL APPROACH FOR PROPOSED STATISTIC

Monitoring of the effectiveness of clinical procedures and physicians' performance has been popularized well over 40 years ago in the medical field (Armitage, 1954 and Bartholomay, 1957). For the health care delivery, patients in hospitals will differ notably in terms of mortality risk. An adjustment for prior risk has to be implemented to ensure that the effect of the true performance of the clinical procedure is not masked by the variability in this prior risk.

Steiner et al. (2000) then proposed the use of a risk-adjusted cumulative sum (CUSUM) chart that accounts for the patient's mortality risk. This risk-adjusted CUSUM chart is formulated based on testing the odds ratio of the mortality of a patient, where H_0 : odds ratio $= Q_0$ versus H_A : odds ratio $= Q_A$. This is equivalent to testing $H_0 : p_0(x_t)/[1 - p_0(x_t)] = Q_0 x_t/(1 - x_t)$ versus $H_A : p_A(x_t)/[1 - p_A(x_t)] = Q_A x_t/(1 - x_t)$ with x_t being the mortality risk for patient t and the mortality rates $p_0(x_t)$ and $p_A(x_t)$ are functions of the mortality risk x_t. For a fixed value of x_t, these mortality rates $p_0(x_t)$ and $p_A(x_t)$ are constants. In essence, a statistic that can be used to compare hospitals could be formulated based on the odds ratio. In order to obtain our proposed statistic, we will implement a two-step procedure. We will first obtain a kernel-based matching estimator $\hat{p}(x_t; h)$ to estimate $p(x_t)$. We then compute $\hat{Q}(x_t; h)$ for each x_t, and obtain our proposed statistic as the median of these computed values.

For the initial step, we will need to apply an algorithm that employs a "distance" threshold to estimate $p(x_t)$ for each x_t. By using kernel-based matching estimators which are commonly used in topological studies, we will form weighted

averages of the post-procedural outcome y_t of all n patients in the sample:

$$\hat{p}(x_t; h) = \frac{\sum_{j=1}^{n} K(\frac{x_t - x_j}{h}) y_j}{\sum_{j=1}^{n} K(\frac{x_t - x_j}{h})}, \tag{4.4}$$

where $K(\cdot)$ is the kernel function which is a probability density function that is symmetric about the origin and integrates to 1 over the domain, and h is a bandwidth parameter which controls the amount of smoothing of the data to obtain the estimate. We have investigated various kernel function $K(\cdot)$ developed in the literature and the Gaussian kernel function with bandwidth $h = 0.9n^{-1/5} \min\{s, IQR/2.68\}$ where IQR is the sample interquartile range and s is the sample standard deviation, proposed by Chen and Kelton (2006) provides satisfactory smoothing performance and emanates $\hat{Q}$ adequately. Details can be found in the Appendix of Chapter 3.

After we have estimated $p(x_t; h)$, we will then compute the estimated odds ratio for each x_t:

$$\hat{Q}(x_t; h) = \frac{\hat{p}(x_t; h)(1 - x_t)}{x_t(1 - \hat{p}(x_t; h))}, \tag{4.5}$$

and since the mortality risk distribution of patients is predominantly skewed, we will obtain the proposed statistic as:

$$\tilde{Q} = median\{\hat{Q}(x_1; h), \hat{Q}(x_2; h), \cdots, \hat{Q}(x_n; h)\}, \tag{4.6}$$

Upon obtaining $\tilde{Q}$ for each hospital, the hospitals are ranked based on the values of their corresponding $\tilde{Q}$, with the smaller values of $\tilde{Q}$ denoting better performances.

SECTION 4. NUMERICAL EXAMPLES

To illustrate the procedure to use our proposed statistic, as well as to compare the performance of this statistic and SMR, some illustrations of examples using real data are shown and their corresponding results are presented. For the first example, the patients underwent cardiac surgery operations in an anonymous hospital in UK and their post-operative outcomes after thirty days were collected. The corresponding mortality risk for each patient was both calculated and authenticated locally at the hospital. The data is stratified based on 3 physicians, which we will identify them as Physicians A, B and C. We calculate the SMR values for Physicians A, B and C to be 1.3270, 1.2013 and 1.1932 respectively. This leads to a conclusion that despite all 3 physicians exhibiting poor performance, the physicians differ in their performances, with Physician C showing the best performance and Physician A showing the worst performance amongst the three. This conjecture could be flawed due to discussed reasons in Section 2.

By investigating the average mortality risk of patients seen by each physician, that for Physicians A, B and C are found to be 0.0343, 0.0426 and 0.0445, with Physician A operating on more patients of lower risks and Physician C operating on more patients of higher risk. As such, the discrepancies in SMR values are possibly due to differences in the mortality risk distributions. Using our proposed statistics in (4.6), the performances between the 3 physicians are found to be similar with $\tilde{Q} = 1.2473$, 1.2433 and 1.2422.

Similarly, for the second example, the patients underwent the same type of surgical operations in another anonymous hospital in UK, and their post-operative and mortality risk were collected. Likewise, the 3 physicians that conducted the surgeries are identified as Physician A, B and C. The calculated SMR values are 1.3242, 1.2343 and 1.1862 respectively. Based on these SMR values, we might be lead to conclude that all 3 physicians exhibit poor performance, with Physician C showing the best performance and Physician A showing the worst performance.

However, the average mortality risk of patients seen by the 3 physicians are 0.0290, 0.0450 and 0.0462 respectively. As discussed earlier, the SMR values will be affected by both the true underlying performances of the physicians and the mortality risk distribution of the patients seen by each physician. Since all 3 SMR values are above 1 and that the average mortality risk of patients seen by the 3 physicians are different, the conclusion drawn earlier might be incorrect. In fact, using our proposed statistics in (4.6) with $\tilde{Q} = 1.1209$, 1.2863 and 1.8188 respectively, we will reach a contradicting conclusion that Physician C actually showed the worst performance while Physician A showed the best performance.

For our last example, we investigate the performances of 5 physicians (identified as A, B, C, D and E) in yet another hospital. The calculated SMR values for the 5 physicians were 1.2220, 0.8273, 0.6833, 0.6300 and 0.4790 respectively. With reference to the SMR values, we might conclude that all 5 physicians exhibit good performance, with Physician E showing the best performance and Physician A showing the worst performance.

However, the average mortality risk of patients seen by the 5 physicians were 0.0472, 0.0605, 0.0563, 0.0390 and 0.0271 respectively. Since the SMR values will be affected by both the true underlying performances of the physicians and the mortality risk distribution of the patients seen by each physician and that 4 of the SMR values are below 1, with the average mortality risk of patients seen by those 4 physicians being different, the conclusion drawn earlier might be incorrect. The only conclusion that we can reach safely will be that Physician A shows the worse performance since it has an SMR of greater than 1. In fact, using our proposed statistics in (4.6) with $\tilde{Q} = 1.3170$, 0.2500, 0.5160 0.6371 and 0.8172 respectively, we will reach the conclusion that the ranking of the Physicians with the best performance to the worst performance should be Physician B, C, D, E and A.

SECTION 5. CONCLUSION

The evolution of the assessment of medical practice has been speeding up tremendously, as seen from recent literature (Werner and Bradlow, 2006, Clarke and Oakley, 2007, Krumholz et al., 2008, Biswas and Kalbfleisch, 2008, Steiner and Jones, 2009). One such widely-used overall quality indicator and measurement tool will be the SMR. The theoretical estimates of the mean and variance for SMR are presented in this paper. In the literature to compare health care providers, there has always been concerns on the non-standardized documentation and coding of patients conditions, as well as on the validity of case mix adjustment methods used to predict mortality rates.

Beyond this methodological bias, we observe that if the odds ratio $Q = 1$,

regardless of whether the mortality risk distributions are different, $E(SMR)$ is always equal to 1. The SMR is robust to differences in the risk distribution. We also observe that if one hospital has a SMR value greater than 1 while another has that smaller than 1, we can compare these two hospitals and conclude that the latter is performing better than the former. However, if both hospitals have SMR values greater than 1 or smaller than 1, we are not able to compare these two hospitals without looking at the risk distribution. Moveover, SMR depend on both the true underlying performances of the hospitals and the mortality risk distribution of the patients seen at each hospital, as seen from our theoretical study.

Some possibly wrong and proper interpretations through the use of SMR in the literature. An alternative statistic for the comparison of hospitals, formulated based on the odds ratio, is proposed. To illustrate the procedure to use our proposed statistic, as well as to compare the performance of this statistic and SMR, some illustrations of examples using real data are shown and their corresponding results are also presented.

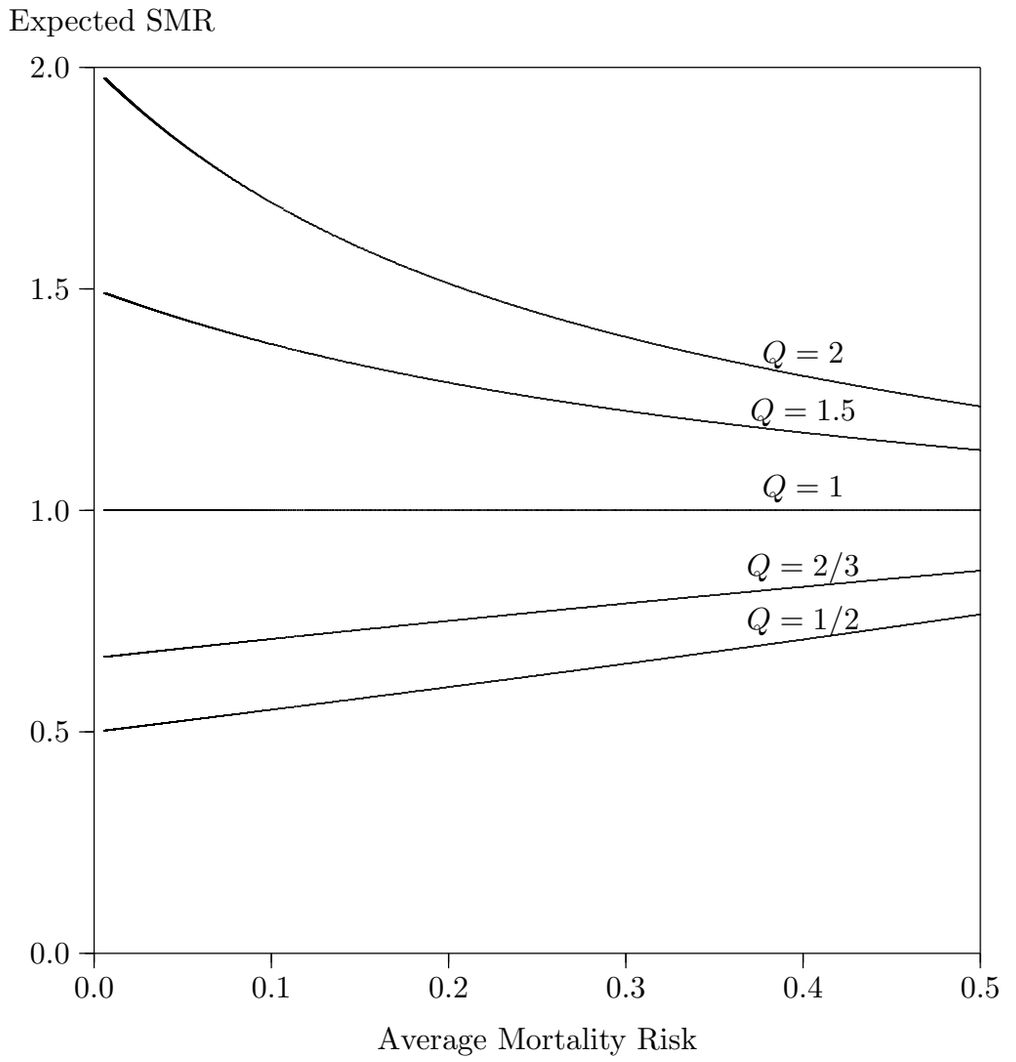

Figure 4.1. Plot of E(SMR) against average mortality risk, with the mortality risk distribution as beta(1,β) and $n = 1000$.

CHAPTER 5. CONCLUSION

We have proposed a joint monitoring scheme for clinical performance and the mortality risk. We have demonstrated that this scheme is easy to implement and is essential to avoid making erroneous inferences on clinical performance when the mortality risk distribution has changed. A new charting procedure to monitor the average mortality risk of patients is also proposed.

Changing the focus from process monitoring, we have proposed a set of risk-adjusted test procedures which alleviate the problem of interpretation through the use of penalty-reward scores. Other risk-adjusted method, specifically the logistic regression is also discussed. However, the use of these test procedures require a model between the mortality rates and mortality risks to be specified. The evaluation of whether a model is wrongly used is not trivial. As such, we have proposed a model-free diagnostic technique to evaluate the effectiveness of the clinical procedure by investigating the mortality rates and resulting odds ratio function against the mortality risks.

Lastly, we presented some facts and myths on the usage of SMR. The various limitations of using the SMR is adequately discussed with various possibly wrong interpretations being highlighted. The theoretical estimates of the mean and variance for SMR are also derived.

For future works, one could develop possibly a distribution-free overall quality indicator that truly adjusts for mortality risks and a criterion to determine the efficiency of quality indicators are. These works can be extended beyond medical field to other fields, as well as accommodate other than Bernoulli outcomes.

BIBLIOGRAPHY

ABRAMOWITZ, M. AND STEGUN, I. A. (1968). Handbook of Mathematical Functions. *Dover Publications Inc., New York*.

ALTMAN, D. G.; GORE, S. M.; GARDNER, M. J. AND POCOCK, S. J. (1983). Statistical Guidelines for Contributors to Medical Journals. *British Medical Journal (Clinical Research Edition)* **286(6376)**, 1489–1493.

AMIN, J.; LAW, M. G.; BARTLETT, M.; KALDOR, J. M. AND DORE, G. J. (2006). Causes of Death After Diagnosis of Hepatitis B or Hepatitis C Infection: A Large Community-Based Linkage Study. *The Lancet* **368**, 938–945.

ARMITAGE, P. (1954). Sequential Tests in Prophylactic and Therapeutic Trials. *Quarterly Journal of Medicine* **23**, 255–274.

BARTHOLOMAY, A. F. (1957). Sequential Probability Ratio Test Applied to the Design of Clinical Experiments. *New England Journal of Medicine* **256(11)**, 498–505.

BBC NEWS (1998). Health Hospitals to get quality control. *News Article, UK*. Available at: http://news.bbc.co.uk/2/hi/health /105787.stm (accessed Mar 2009).

BERGER, A.; SADOSKY, A.; DUKES, E.; MARTIN, S.; EDELSBERG, J. AND OSTER, G. (2008). Characteristics and Patterns of Healthcare Utilization of Patients with Fibromyalgia in General Practitioner Settings in Germany. *Current Medical Research and Opinion* **24(9)**, 2489–2499.

BERRY, W. D. AND FELDMAN, S. (1985). Multiple Regression in Practice. *Sage Publications, United States*.

BISWAS, P. AND KALBFLEISCH, J. D. (2008). A Risk-Adjusted CUSUM in Continuous Time Based on the Cox Model. *Statistics in Medicine* **27**, 3382–3406.

BOCCASANTA, P.; VENTURI, M.; SPENNACCHIO, M.; BUONAGUIDI, A.; AIROLDI, A. AND ROVIARO, G. (2009). Prospective Clinical and Functional Results of Combined Rectal and Urogynecologic Surgery in Complex Pelvic Floor Disorders. *The American Journal of Surgery* **199(2)**, 144–153.

BRESLOW, N. E. AND DAY, N. E. (1987). Statistical Methods in Cancer Research: The Design and Analysis of Cohort Studies. *IARC Scientific Publications: France*.

BRISTOL ROYAL INFIRMARY INQUIRY (2001). Learning from Bristol: the Report of the Inquiry into Childrens Heart Surgery at the Bristol Royal Infirmary 1984 ?1995. *Report, UK*. Available at: http://www.bristol-inquiry.org.uk/ (accessed Mar 2009).

BURKOM, H. (2006). Discussion - The Use of Control Charts in Health-Care and Public Health Surveillance. *Journal of Quality Technology* **38(2)**, 127–132.

CARDOZO, L. D.; STANTON, S. L.; ROBINSON, H. AND HOLE, D. (1980). Evaluation of Flurbiprofen in Detrusor Instability. *British Medical Journal* **280**, 281–282.

CHAMBLER, A. F. AND EMERY, R. J. (1997). Lord Moynihan Cuts Codman into Audit. *Annals of the Royal College of Surgeons of England* **79**, 174–176.

CHEN, E. J. AND KELTON, W. D. (2006). Empirical Evaluation of Data-Based Density Estimation. *Proceedings of the 38th conference on Winter Simulation*,

333–341.

CHEN, R. (1978). A Surveillance System for Congenital Malformations. *Journal of American Statistical Association* **73**, 323–327.

CHEN, R. (1996). Exploratory Analysis as a Sequel to Suspected Increased Rate of Cancer in a Small Residential or Workplace Community. *Statistics in Medicine* **15**, 807–816.

CHEN, Y. Y.; CONNORS, A. F. AND GARLAND, A. (2008). Effect of Decisions to Withhold Life Support on Prolonged Survival. *Chest* **133(6)**, 1312–1318.

CLARKE, S. AND OAKLEY, J. (2007). Informed Consent and Clinician Accountability: the Ethics of Report Cards on Surgeon Performance. *Cambridge University Press, Cambridge.*

CONSORTIUM OF CHIEF QUALITY OFFICERS (2009). Using Hospital Standardized Mortality Ratios for Public Reporting: A Comment by the Consortium of Chief Quality Officers: Consortium of Chief Quality Officers. *American Journal of Medical Quality* **24(2)**, 164–165.

COOK, D. A.; DUKE, G.; HART, G. K.; PILCHER, D. AND MULLANY, D. (2008). Review of the Application of Risk-Adjusted Charts to Analyse Mortality Outcomes in Critical Care. *Critical Care Resuscitation* **10(3)**, 239–251.

COOK, D. A.; STEINER, S. H.; FAREWELL, V. T. AND MORTON, A. P. (2003). Monitoring the Evolutionary Process of Quality: Risk Adjusted Charting to Track Outcomes in Intensive Care. *Critial Care Medicine* **31**, 1676–1682.

COORY, M.; DUCKETT, S. AND SKETCHER-BAKER, K. (2008). Using Con-

trol Charts to Monitor Quality of Hospital Care with Administrative Data. *Journal for Quality in Health Care* **20(1)**, 31–39.

COURT. B. V. AND CHENG. K. K. (1995). Pros and Cons of Standardised Mortality Ratios. *The Lancet* **346**, 1432.

DEHEUVELS, P. (1977). Estimation Non Parametrique de la densite par histogrammes generalises. *Rev. Stat. Appi.* **25(3)**, 5–43.

DICKERSON, J.; HINGORANI, A.; ASHBY, M.; PALMER, C. AND BROWN, M. (1999). Optimization of Antihypertensive Treatment by Crossover Rotation of Four Major Classes. *The Lancet* **353(9169)**, 2008–2013.

DOOLEY, W. C.; LJUNG, B. M.; VERONESI, U.; CAZZANIGA, M.; ELLEDGE, R. M.; O'SHAUGH-NESSY, J. A.; KUERER, H. M.; HUNG, D. T.; KHAN, S. A.; PHILLIPS, R. F.; GANZ, P. A.; EUHUS, D. M.; ESSERMAN, L. J.; HAFFTY, B. G.; KING, B. L.; KELLEY, M. C.; ANDERSON, M. M.; SCHMIT, P. J.; CLARK, R. R.; KASS, F. C.; ANDERSON, B. O.; TROYAN, S. L.; ARIAS, R. D.; QUIRING, J. N.; LOVE, S. M.; PAGE, D. L. AND KING, E. B. (2001). Ductal Lavage for Detection of Cellular Atypia in Women at High Risk for Breast Cancer. *Journal of the National Cancer Institute* **93(21)**, 1624–1632.

DORSCH, M. F.; LAWRANCE, R. A.; SAPSFORD, R. J.; OLDHAM, J.; GREENWOOD, D. C.; JACKSON, B. M.; MORRELL, C.; BALL, S. G.; ROBINSON, M. B.; HALL, A. S. AND THE EMMACE STUDY GROUP. (2001). A Simple Benchmark for Evaluating Quality of Care of Patients Following Acute Myocar-

dial Infarction. *Heart* **86**, 150–154.

DR FOSTER INTELLIGENCE (2007). Hospital Guide. *Hospital Guide, UK*. Available at: http://www.drfoster.co.uk/hospitalguide (accessed Jun 2010).

DUPRET, G. AND KODA, M. (2001). Bootstrap Re-Sampling for Unbalanced Data in Supervised Learning. *European Journal of Operational Research* **134(1)**, 141–156.

EGGER, M.; ZELLWEGER-ZAHNER, T.; SCHNEIDER, M.; JUNKER, C.; LENGELER, C. AND ANTES, G. (1997). Language Bias in Randomized Controlled Trials published in English and German. *The Lancet* **350(9074)**, 326–329.

FRISEN, M. AND DE MARE, J. (1991). Optimal Surveillance. *Biometrika* **78**, 271–280.

FRISEN, M. (1992). Evaluations of Methods for Statistical Surveillance. *Statistics in Medicine* **11**, 1489–1502.

GALLUS, G.; MANDELLI, C.; MARCHI, M. AND RAAELLI, G. (1986). On Surveillance Methods for Congenital Malformations. *Statistics in Medicine* **5**, 565–571.

GAN. F. F. AND TAN, T. (2010). Risk-Adjusted Number-Between Failures Charting Procedures for Monitoring a Patient Care Process for Acute Myocardial Infarctions. *Health Care Management Science*, DOI 10.1007/s10729-010-9125-8.

GLANCE, L. G.; OSLER, T. AND SHINOZAKI, T. (2000). Effect of Varying the Case Mix on the Standardized Mortality Ratio and W Statistic: A Simulation

Study. *Chest* **117**, 1112–1117.

GORDON, R. R. (1995). Mortality and Social Deprivation. *The Lancet* **345**, 1640.

GREINACHER, A.; AMIRAL, J. J.; DUMMEL, V.; VISSAC, A.; KIEFEL, V.; MUELLER-ECKHARDT, C. (1994). Laboratory Diagnosis of Heparin-Associated Thrombocytopenia and Comparison of Platelet Aggregation Test, Heparin-Induced Platelet Activation Test, and Platelet Factor 4/Heparin Enzyme-Linked Immunosorbent Assay. *Transfusion* **34(5)**, 381–385.

GRIGG, O. A. AND FAREWELL, V. T. (2004). A Risk-Adjusted Sets Method for Monitoring Adverse Medical Outcomes. *Statistics in Medicine* **23**, 1593–1602.

GRIGG, O. AND SPIEGELHALTER D. (2007). A Simple Risk-Adjusted Exponentially Weighted Moving Average. *Journal of the American Statistical Association* **102(477)**, 140–152.

HAKES, J. AND VISCUSI, W. K. (1997). Mortality Risk Perceptions: a Bayesian Reassessment. *Journal of Risk and Uncertainty* **15**, 135–150.

HACKBUSCH, W. (1995). Integral Equation: Theory and Numerical Treatment. *International Series of Numerical Mathematics* **120**, Birkhäuser Verlag, Basel, Boston, Berlin.

HOSPITAL QUALITY ALLIANCE. Hospital Quality Reporting in the United States: A Cost Analysis for the Hospital Quality Alliance", Report, United States, 2006. Available at:

www.hospitalqualityalliance.org (accessed Dec 2009).

HOWELL, J. (1995). Standardised Mortality Ratios. *The Lancet* **346**, 904.

HUBER-WAGNER, S.; LEFERING, R.; QVICK, L.; KORNER, M.; KAY, M. V.; PFEIFER, K.; REISER, M.; MUTSCHLER, W. AND KANZ, K. (2009). Effect of Whole-Body CT During Trauma Resuscitation on Survival: A Retrospective, Multicentre Study. *The Lancet* **373**, 1455–1461.

IEZZONI, L. I. (1997). The Risks of Risk Adjustment. *Journal of American Medical Association*, 278(19), 1600–1607.

ILLSLEY, R. (1995). Mortality Trends in UK. *The Lancet* **346**, 313–314.

JARMAN, B.; PIETER, D.; VAN DER VEEN, A. A.; KOOL, R. B.; AYLIN, P.; BOTTLE, A.; WESTERT, G. P. AND JONES, S. (2010). The Hospital Standardised Mortality Ratio: A Powerful Tool for Dutch Hospitals to Assess Their Quality of Care? *Quality and Safety in Health Care* **19(1)**, 9–13.

JOHNSTON, S. L.; PATTEMORE, P. K.; SANDERSON, G.; SMITH, S.; LAMPE, F.; JOSEPHS, L.; SYMINGTON, P.; O'TOOLE, S.; MYINT, S. H.; TYRRELL, D. A. J. AND HOLGATE, S. T. (1995). Community Study of Role of Viral Infections in Exacerbations of Asthma in 9-11 Year Old Children. *British Medical Journal* **310**, 1225–1229.

JOHNSON, C. J.; MORTON, A.; ROBINSON, M. B. AND HALL, A. (2005). Real-Time Monitoring of Coronary Care Mortality: A Comparison and Combination of Two Monitoring Tools. *International Journal of Cardiology* **100**, 301–307.

JONES, J. M.; REDMOND, A. D. AND TEMPLETON, J. (1995). Uses and

Abuses of Statistical Models for Evaluating Trauma Care *Journal of Trauma* **38**, 89–93.

JULIOUS, S. A.; NICHOLL, J. AND GEORGE, S. (2001). Why do We Continue to Use Standardized Mortality Ratios for Small Area Comparisons? *Journal of Public Health Medicine* **23**, 40–46.

KASKA, S. C. AND WEINSTEIN, J. N. (1998). Historical Perspective: Ernest Amory Cod- man, 1869-1940. A Pioneeer of Evidence-Based Medicine: The End Result Idea. *Spine* **23**, 629–633.

KENETT, R. AND POLLAK, M. (1983). On Sequential Detection of a Shift in the Probability of a Rare Event. *Journal of American Statistical Association* **78**, 389–395.

KNOTH, S. (2005). Accurate ARL Computation for EWMA-S^2 Control Charts. *Statistics and Computing* **15**, 341–352.

KNOTH, S. (2007). Accurate ARL Computation for EWMA Control Charts Monitoring Normal Mean and Variance Simultaneously. *Sequential Analysis* **26**, 251–263.

KOOPMANS, K. P.; NEELS, O. C.; KEMA, I. P.; ELSINGA, P. H.; SLUITER, W. J.; VANGHILLEWE, K.; BROUWERS, A. H.; JAGER, P. L. AND VRIES, E. G. E. (2008). Improved Staging of Patients with Carcinoid and Islet Cell Tumors with ^{18}F-Dihydroxy-Phenyl-Alanine and ^{11}C-5-Hydroxy-Tryptophan Positron Emission Tomography. *Journal of Clinical Oncology* **26(9)**, 1489–1495.

KRUMHOLZ, H. M.; KEENAN, P. S.; BRUSH, J. E.; BUFALINO, V. J.; CHERNEW,

M. E.; EPSTEIN, A. J.; HEIDENREICH, P. A.; HO, V.; MASOUDI, F. A.; MATCHAR, D. B.; NORMAND, S. T.; RUMSFELD, J. S.; SCHUUR, J. D.; SMITH, J. S. C.; SPERTUS, J. A. AND WALSH, M. N. (2008). Standards for Measures Used for Public Reporting of Efficiency in Health Care. *Circulation* **118**, 1885–1893.

KRUMHOLZ, H. M.; WANG, Y.; CHEN, J.; DRYE, E. E.; SPERTUS, J. A.; ROSS, J. S.; CURTIS, J. P.; NALLAMOTHU, B. K.; LICHTMAN, J. H.; HAVRANEK, E. P.; MASOUDI, F. A.; RADFORD, M. J.;HAN, L. F.; RAPP, M. T.; STRAUBE, B. M. AND NORMAND S. T. (2009). Reduction in Acute Myocardial Infarction Mortality in the United States: Risk-Standardized Mortality Rates from 1995-2006. *Journal of American Medical Association* **302(7)**, 767–773.

KUIPERS, E. J.; LUNDELL, L.; KLINKENBERG-KNOL, E. C.; HAVU, N.; FESTEN, H. P. M.; LIEDMAN, B.; LAMERS, C. B. H. W.; JANSEN, J. B. M. J.; DALENBACK, J.; SNEL, P.; NELIS, G. F. AND MEUWISSEN, S. G. M. (1996). Atrophic Gastritis and Helicobacter Pylori Infection in Patients with Reflux Esophagitis Treated with Omeprazole or Fundoplication. *The New England Journal of Medicine* **334(16)**, 1018–1022.

LILFORD, R.; MOHAMMED, M. A.; SPIEGELHALTER, D. AND THOMSON, R. (2004). Use and Misuse of Process and Outcome Data in Managing Performance of Acute Medical Care: Avoiding Institutional Stigma. *The Lancet* **363**, 1147–1154.

LITTELL, R. C.; STROUP, W. W. AND FREUND, R. J. (2002). SAS for Linear Models). *SAS Publishing, United States*.

LIU, Y.; CHAWLA, N.; HARPER, M.; SHRIBERG, E. AND STOLCKE, A. (2006). A Study in Machine Learning for Imbalanced Data for Sentence Boundary Detection in Speech. *Computer Speech and Language* **20(4)**, 468–494.

LOVEGROVE, J.; VALENCIA, O.; TREASURE, T.; SHERLAW-JOHNSON, C. AND GALLIVAN, S. (1997). "Monitoring the Results of Cardiac Surgery by Variable-Life-Adjusted Display. *The Lancet* **350**, 1128–1130.

LOVEGROVE, J.; SHERLAW-JOHNSON, C.; VALENCIA, O. AND GALLIVAN, S. (1999). Monitoring the Performance of Cardiac Surgeons. *Journal of the Operational Research Society* **50**, 685–689.

LOVEGROVE, J.; VALENCIA, O.; TREASURE, T.; SHERLAW-JOHNSON, C. AND GALLIVAN, S. (1997). Monitoring the Results of Cardiac Surgery by Variable-Life-Adjusted Display. *The Lancet* **350**, 1128–1130.

MATHENY, M. L.; OHNO-MACHADO, L. AND RESNIC, F. (2007). Risk-Adjusted Sequential Probability Ratio Test Control Chart Methods for Monitoring Operator and Institutional Mortality Rates in Interventional Cardiology. *American Heart Journal* **155**, 114-C120.

MAXWELL, E. (1970). Comparing the Classification of Subjects by Two Independent Judges. *The British Journal of Psychiatry* **116**, 651–655.

MCCULLAGH, P. AND NELDER, J. (1989). Generalized Linear Models (2nd Edition). *Chapman & Hall/CRC, United Kingdom*.

MCNEMAR, Q. (1947). Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **12(2)**, 153–157.

MOHAMMED, M. A.; DEEKS , J. J.; GIRLING, A.; RUDGE, G.; CARMALT, M.; STEVENS, A. J. AND LILFORD, R. (2009). Evidence of Methodological Bias in Hospital Standardised Mortality Ratios: Retrospective Database Study of English Hospitals. *British Medical Journal* **338**, 1–8.

MOITRA, S. D. (1990). Skewness and the Beta Distribution. *The Journal of the Operational Research Society* **41**, 953–961.

MOORE, L.; HANLEY, J. A.; TURGEON, A. F.; LAVOIE, A. AND ERIC, B. (2010). A New Method for Evaluating Trauma Centre Outcome Performance. *Annals of Surgery* **251(5)**, 952–958.

MORGAN, M. G. AND HENRION, M. (1990). Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. *Cambridge University Press, New York*.

MOUSTAKIDES, G. V. (1986). Optimal Stopping Times for Detecting Changes in Distributions. *The Annals of Statistics* **14**, 1379–1387.

NASHEF, S. A. M.; ROQUES, F.; MICHEL, E. G., LEMESHOW, S.; SALAMON, R. AND THE EUROSCORE STUDY GROUP (1999). European System for Cardiac Operative Risk Evaluation (EuroSCORE). *European Journal of Cardio-Thoracic Surgery* **16(1)**, 9–13.

NATIONAL ADULT CARDIAC SURGICAL DATABASE REPORT) (2001). National Adult Cardiac Surgical Database Report 1999 - 2000. *Report, UK*. Avail-

able at: http://www.scts.org /doc/5483 (accessed Mar 2009).

NEW YORK STATE DEPARTMENT OF HEALTH. New York State Hospital Profile, Report, United States, 2008. Available at: *hospitals.nyhealth.gov* (accessed Dec 2009).

NEYMAN, J. AND PEARSON, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A* **231**, 289–337.

NOVICK, R. J.; FOX, S. A.; STITT, L. W. AND FORBES, T. L. (2006). Direct Comparison of Risk-Adjusted and Non-Risk-Adjusted CUSUM Analysis of Coronary Artery Bypass Surgery Outcomes. *The Journal of Thoracic and Cardiovascular Surgery* **132**, 386–391.

PAGE, E. S. (1954). Continuous Inspection Schemes. *Biometrika* **41**, 100?115.

PAPADOPOULOS, F. C.; EKBOM, A.; BRANDT, L. AND EKSELIUS, L. (2009). Excess Mortality, Causes of Death and Prognostic Factors in Anorexia Nervosa. *The British Journal of Psychiatry* **194**, 10–17.

PARK, B. AND MARRON, J .S. (1990). Comparison of Data-Driven Bandwidth Selectors. *Journal of the American Statistical Association* **85**, 66–72.

PARSONNET, V.; BERNSTEIN, A. D. AND GERA, M. (1996). Clinical Usefulness of Risk-Stratified Outcome Analysis in Cardiac Surgery in New Jersey. *Annals of Thoracic Surgery* **61**, S8–S11.

PARSONNET, V.; DEAN, D. AND BERNSTEIN, A. D. (1989). A Method of Uniform Stratification of Risk for Evaluating the Results of Surgery in Acquired

Adult Heart Disease. *Circulation* **79(6:2)**, I3–I2.

POLONIECKI, J.; VALENCIA, O. AND LITTLEJOHNS, P. (1998). Cumulative Risk Adjusted Mortality Chart for Detecting Changes in Death Rate: Observational Study of Heart Surgery. *British Medical Journal* **316**, 1697–1700.

QUIGLEY, D. D.; ELLIOTT, M. N.; HAYS, R. D.; KLEIN, D. J. AND FARLEY, D. O. (2008). Bridging from the Picker Hospital Survey to the CAHPS(R) Hospital Survey. *Medical Care* **46(7)**, 654–661.

ROGERS, C. A.; REEVES, B. C.; CAPUTO, M.; GANESH, J. S.; BOSNER, R. S. AND ANGELINI, G. D. (2004). Control Chart Methods for Monitoring Cardiac Surgical Performance and Their Interpretation. *The Journal of Thoracic and Cardiovascular Surgery* **128**, 811–819.

ROSSI, G.; LAMPUGNANI, L. AND MARCHI, M. (1999). An Approximate CUSUM Procedure for Surveillance of Health Events. *Statistics in Medicine* **18**, 2111–2122.

SCHATZKIN, A.; CONNOR, R. J.; TAYLOR, P. R. AND BUNNAG, B. (1987). Comparing New and Old Screening Tests when a Reference Procedure cannot be Performed on All Screenees: Example of Automated Cytometry for Early Detection of Cervical Cancer. *American Journal of Epidemiology* **125(4)**, 672–678.

SCHWARTZ, P. J.; LOCATI, E. H.; MOSS, A. J.; CRAMPTON, R. S.; TRAZZI, R. AND RUBERTI, U. (1991). Left Cardiac Sympathetic Denervation in the Therapy of Congenital Long QT Syndrome: A Worldwide Report. *Circulation*

84, 503–511.

SCOTT, R.; BESAG, F. AND NEVILLE, B. (1999). Buccal Midazolam and Rectal Diazepam for Treatment of Prolonged Seizures in Childhood and Adolescence: A Randomized Trial. *The Lancet* **353(9153)**, 623–626.

SEEMAN, E.; MELTON, L.; FALLON, W. O. AND RIGGS, B. (1983). Risk Factors for Spinal Osteoporosis in Men. *The American Journal of Medicine* **75(6)**, 977–983.

SHEATHER, S. J. AND JONES, M. C. (1991). A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation. *Journal of the Royal Statistical Society, Series B* **53**, 683–690.

SHERLAW-JOHNSON, C.; MORTON, A.; ROBINSON, M. B. AND HALL, A. (2005). Real-Time Monitoring of Coronary Care Mortality: A Comparison and Combination of Two Monitoring Tools. *International Journal of Cardiology* **100**, 301–307.

SHERLAW-JOHNSON, C.; WILSON, P. AND GALLIVAN, S. (2007). The Development and Use of Tools for Monitoring the Occurrence of Surgical Wound Infections. *Journal of the Operational Research Society* **58**, 228–234.

SILVERMAN, B. W. (1981). Using Kernel Density Estimates to Investigate Multimodality. *Journal of the Royal Statistical Society, Series B* **43**, 97–99.

SILVERMAN, B. W. (1986). Density Estimation for Statistics and Data Analysis. *Chapman and Hall, London.*

SPIEGELHALTER, D. (1999). Surgical Audit: Statistical Lessons from Nightin-

gal and Codman. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **162**, 45–58.

SPIEGELHALTER, D. (2004). Monitoring Clinical Performance: A Commentary. *Journal of Thoracic and Cardiovascular Surgery* **128(6)**, 820–822.

SPIEGELHALTER, D.; GRIGG, O.; KINSMAN, R. AND TREASURE, T. (2003). Risk-Adjusted Sequential Probability Ratio Tests: Applications to Bristol, Shipman and Adult Cardiac Surgery. *International Journal for Quality in Health Care* **15**, 7–13.

SPIEGELHALTER, D. J.; GRIGG, O. A.; KINSMAN, R. AND TREASURE, T. (2003). Sequential Probability Ratio Tests (SPRTs) for Monitoring Risk-Adjusted Outcomes. *International Journal for Quality in Health Care* **15**, 1–7.

STEINER, S.; COOK, R. AND FAREWELL, V. (1999). Monitoring Paired Binary Surgical Outcomes Using Cumulative Sum Charts. *Statistics in Medicine 1999* **18(1)**, 69–86.

STEINER, S. H. (2006). Discussion - The Use of Control Charts in Health-Care and Public Health Surveillance. *Journal of Quality Technology* **38(2)**, 111–112.

STEINER, S. H.; COOK, R. J.; FAREWELL, V. T. AND TREASURE, T. (2000). Monitoring Surgical Performances Using Risk-Adjusted Cumulative Sum Charts. *Biostatistics* **1**, 441–452.

STEINER, S. H. AND JONES, M. (2009). Risk-Adjusted Survival Time Monitoring with an Updating Exponentially Weighted Moving Average (EWMA) Control Chart. *Statistics in Medicine* **29(4)**, 444–454.

UHLMANN, R. F.; PEARHMAN, R. A. AND CAIN, K. C. (1988). Physicians' and Spouses' Predictions of Elderly Patients' Resuscitation Preferences. *Journal of Gerontology* **43(5)**, 115–121.

WALD, A. (1947). Sequential Analysis. *John Wiley, New York*.

WERNER, R. M. AND BRADLOW, E. T. (2006). Relationship Between Medicares Hospital Compare Performance Measures and Mortality Rates. *Journal of the American Medical Association* **296(22)**, 2694–2702.

WOODALL W. H. (2000). Controversies and Contradictions in Statistical Process Control. *Journal of Quality Technology* **32**, 341-C350.

YAN, T.; YANG, Y. N.; CHENG, X.; DEANGELIS, M. M., HOH, J. AND ZHANG, H. (2008). Genotypic Association Analysis Using Discordant-Relative-Pairs. *Annals of Human Genetics* **73(1)**, 84–94.

APPENDIX A: RISK-ADJUSTED CUSUM CHART TO MONITOR CLINICAL PERFORMANCES

This monitoring scheme will be conducted sequentially after the clinical procedure is implemented on each patient. If we let y to be the post-procedural outcome for a patient, it corresponds to one of two possible outcomes (success or failure). We assume y_t is the outcome for patient t. (1 if there is a mortality or 0 if a patient survives after implementation of the clinical procedure). Notice that the outcome of the clinical procedure may not be observed immediately after its implementation, with one example being that for cardiac operations in which the outcome of mortality is usually determined within 30 days from surgery. We obtained the following probability function of y_t, $f(y_t|p) = p^{y_t}[1-p]^{1-y_t}$, where p is the mortality rate.

The risk-adjusted CUSUM chart is formulated based on testing the odds ratio of the mortality of a patient, where H_0 : odds ratio $= Q_0$ versus H_A : odds ratio $= Q_A$. This is equivalent to testing $H_0 : p_0(x_t)/[1-p_0(x_t)] = Q_0 x_t/(1-x_t)$ versus $H_A : p_A(x_t)/[1-p_A(x_t)] = Q_A x_t/(1-x_t)$ with x_t being the mortality risk for patient t and the mortality rates $p_0(x_t)$ and $p_A(x_t)$ are functions of the mortality risk x_t. For a fixed value of x_t, these mortality rates $p_0(x_t)$ and $p_A(x_t)$ are constants. The possible log-likelihood ratio score for patient t is:

$$W_t|x_t = \begin{cases} \log\left\{\dfrac{(1-x_t+Q_0x_t)Q_A}{(1-x_t+Q_Ax_t)Q_0}\right\}, & \text{if } y_t = 1, \\ \log\left\{\dfrac{1-x_t+Q_0x_t}{1-x_t+Q_Ax_t}\right\}, & \text{if } y_t = 0. \end{cases} \tag{A1}$$

The CUSUM test statistic S_t, calculated using data up to and including patient t, is $S_t = \max\{0, S_{t-1} + W_t\}$ where $S_0 = 0$, with a lower holding barrier at 0

which is due to the fact that the acceptance of H_0 is not of primary interest under practicality of continual monitoring.

APPENDIX B: CUMULATIVE SUM CHART TO MONITOR MORTALITY RISK DISTRIBUTION

Since the mortality risk x_t is between 0 and 1, and from previous studies of the mortality risk distribution, its theoretical model distribution may be modeled as beta(α,β). Morgan and Henrion (1990) and Moitra (1990) pointed out that the use of beta distributions in modeling data in the form of proportions is due to its large variety of shapes. Moreover, Hakes and Viscusi (1997) also discussed that since the beta distribution is flexible and can assume a vast variety of skewed and symmetric shapes, its use is not notably restrictive. However, it is also noted that our proposed charting procedure is not confined to the use of only beta distribution. Through slight modifications, this procedure can be used for other distributions for the risk.

On another note, past literature has shown that the average mortality risk of patients undergoing cardiac surgery has been predominantly increasing over the years. Parsonnet, Bernstein and Gera (1996) demonstrated that when the Parsonnet model was applied to the patients of the Beth Israel Hospital in Newark, the average mortality risk of patients had progressively increased by 47.7% in 1994, as compared to that of 6.5 in 1988. The National Adult Cardiac Surgical Database Report (2001) also showed similar trends when the average mortality risk of patients in 1999 had increased by 20% over 3 years. Since much literature has placed great emphasis on the average mortality risk, we propose a cumulative sum chart to specifically monitor the average mortality risk. We re-parameterize the beta distribution so as to obtain a distribution with one of the parameter as

the average, $\mu = \alpha/(\alpha+\beta)$. As a result, we have the following probability function of X_t as:

$$f(X_t;\mu,\beta) = \frac{(1-X_t)^{(\beta-1)} X_t^{\frac{\mu\beta}{1-\mu}-1}}{B(\frac{\mu\beta}{1-\mu},\beta)}, \qquad (A2)$$

where $B(\frac{\mu\beta}{1-\mu},\beta) = \int_0^1 t^{\frac{\mu\beta}{1-\mu}-1}(1-t)^{\beta-1}dt$, is the beta function. As either of the two parameters μ and β of the above probability distribution could change, the CUSUM chart to monitor the average mortality risk will require β to be set as a constant. The chart is then formulated based on testing the average mortality risk, where $H_0 : \mu = \mu_0$ versus $H_A : \mu = \mu_1$ where μ_0 is the estimated average mortality risk from Phase I analysis of historical data and μ_1 is the shifted average mortality risk. This is equivalent to testing $H_0 : f(X;\mu,\beta) = f(X;\mu_0,\beta)$ versus $H_1 : f(X;\mu,\beta) = f(X;\mu_1,\beta)$ for all $X \in [0,1]$. The plotting statistic of the CUSUM chart for patient t can be derived from the sequential probability ratio test (SPRT) procedure proposed by Wald (1947) and if distribution of the mortality risk is modeled by a beta distribution, the probability function of X_t is given by (A2) and the resulting plotting statistic is:

$$Z_t = \log\left\{\frac{f(X_t;\mu_1,\beta)}{f(X_t;\mu_0,\beta)}\right\}, \qquad (A3)$$

and it can be written as

$$W_t = \frac{Z_t(1-\mu_0)(1-\mu_1)}{\beta(\mu_1-\mu 0)} = \log(X_t) + \frac{(1-\mu_0)(1-\mu_1)}{\beta(\mu_1-\mu 0)} \log\left\{\frac{B(\frac{\mu_0\beta}{1-\mu_0},\beta)}{B(\frac{\mu_1\beta}{1-\mu_1},\beta)}\right\}. \qquad (A4)$$

For other distributions for the mortality risks, one will just need to derive, using the corresponding $f(X_t)$ in (A3).

For the detection of an upward shift in the average mortality risk (that is, $\mu_1 > \mu_0$) and that of a downward shift in the average mortality risk (that is,

$\mu_1 < \mu_0$), the CUSUM test statistics, calculated using data up to and including patient t, is:

$$S_t^+ = \max\{0, S_{t-1}^+ + W_t\}, \tag{A5}$$

$$S_t^- = \max\{0, S_{t-1}^- - W_t\}, \tag{A6}$$

respectively, where $S_0^+ = S_0^- = 0$, with a lower holding barrier at 0 which is due to the fact that the acceptance of H_0 is not of primary interest under practicality of continual monitoring.

APPENDIX C: COLLOCATION METHOD

The collocation method is one of the most recent methods proposed to compute the ARL. Knoth (2005) demonstrated that this method is accurate in computing the ARL when the support is not the entire line. For this method, we consider a CUSUM chart obtained by plotting $S_t = \max(0, S_{t-1} + W_t)$ against the patient number t. Let $L(s_0)$ denote the ARL of the CUSUM chart that starts at $S_0 = s_0$, then

$$L(s_0) = 1 + L(0)P(W_t \leq -s_0; \theta, d) + \int_0^h L(x) f_W(x - s_0; \theta, d) dx \qquad (A7)$$

where θ is the parameter that defines the probability distribution function f_W, and d is the parameter of interest investigated by the CUSUM chart.

The collocation method is to approximate $L(s_0)$ by $\sum_{j=1}^{N} c_j T_j(s_0)$, where $T_j(.)$ is a set of N independent interpolating functions, and c_j's are the unknown constants. To solve for c_j's, we have to choose a set of N nodes in the domain $[0, h]$, then solve the resulting system of linear equations, as discussed in Hackbusch (1995). According to Knoth (2005), the Chebychev polynomials $T_j(z) = \cos(j \arccos(z))$, $j = 0, 1, ..., N-1$, $z \in [-1, 1]$ provide stable numerical quadratures, the corresponding nodes are called Chebychev nodes: $z_i = \cos(\frac{(2i-1)\pi}{2N})$, $i = 1, 2, ..., N$ and $z_i \in [-1, 1]$.

According to (A7), we consider for all $j = 1, 2, ..., N$ Chebychev polynomials in $[0, h]$: $T_j(z) = \cos[(j-1) \arccos(\frac{2z-h}{h})]$ and for all $i = 1, 2, ..., N$ Chebychev nodes in $[0, h]$: $z_i = \frac{h}{2}[1 + \cos(\frac{(2i-1)\pi}{2N})]$. As W_t has an upper support (u) and lower support (l), for each z_i, we change the interval $[0, h]$ to $[l^*, u^*]$, where $l^* = 0$

if $0 \geq l + z_i$ and $l^* = l + z_i$ if $0 < l + z_i$, $u^* = h$ if $h \leq u + z_i$ and $u^* = u + z_i$ if $h > u + z_i$. c_j's can then be solved using the following system of linear equations:

$$\sum_{j=1}^{N} c_j T_j(z_i) = 1 + P(W_t \leq -z_i; \theta, d) \sum_{j=1}^{N} c_j T_j(0) + \sum_{j=1}^{N} c_j \int_{l^*}^{u^*} T_j(x) f_W(x - z_i; \theta, d) dx, \qquad (A8)$$

$i = 1, 2, ..., N$. The integral on the right-hand side can be determined using the Gauss-Legendre quadratures (Abramowitz and Stegun, 1968).

As discussed earlier, in order to compute the ARL accurately, the collocation method is adapted using the distribution function of W_t. Suppose the risk-adjusted CUSUM chart is formulated based on testing the odds ratio of the mortality of a patient, where H_0 : odds ratio $= Q_0$ versus H_A : odds ratio $= Q_A$. Usually $Q_0 = 1$, as the estimated risk x_t is based on the current conditions before taking into account the effect of the true performance of the clinical procedure. Consider the log-likelihood ratio score W_t for patient t given in (A1), we can obtain the probability distribution function of W using a conditioning approach (see Ross, 2006, page 376 for examples) as:

$$f_W(w; \theta, d) = \begin{cases} \frac{Q(e^w - Q_A)}{Q_A - e^w Q_A + Q(e^w - Q_A)} \frac{Q_A}{e^w(Q_A - 1)} f_X(\frac{e^w - Q_A}{e^w(1 - Q_A)}; \theta), & log(Q_A) > w \geq 0; \\ \frac{1 - Q_A e^w}{1 - Q_A e^w + Q(e^w - 1)} \frac{1}{e^w(Q_A - 1)} f_X(\frac{e^w - 1}{e^w(1 - Q_A)}; \theta), & -log(Q_A) < w < 0, \end{cases} \qquad (A9)$$

for $Q_A > 1$ and

$$f_W(w; \theta, d) = \begin{cases} \frac{1 - e^w Q_A}{1 - e^w Q_A + Q(e^w - 1)} \frac{1}{e^w(1 - Q_A)} f_X(\frac{e^w - 1}{e^w(1 - Q_A)}; \theta), & -log(Q_A) > w \geq 0; \\ \frac{Q(e^w - Q_A)}{Q_A - e^w Q_A + Q(e^w - Q_A)} \frac{Q_A}{e^w(1 - Q_A)} f_X(\frac{e^w - Q_A}{e^w(1 - Q_A)}; \theta), & log(Q_A) < w < 0, \end{cases} \qquad (A10)$$

for $Q_A < 1$, where Q is the actual odds ratio and f_X is the probability function of the mortality risk, with an example given by (A2).

However in some scenarios, the odds ratio might not be the same across all levels of mortality risks, but instead it is more likely to be a linear function of the mortality risk x_t. As such, in order to be able to compute the ARL accurately, the distribution function of W_t under $Q_0 = 1$ and $Q_A(x_t) = (b-a)x_t + a$ is also derived. We define:

for $y_t = 1$,

$$\frac{Q_A(x_t)}{1 - x_t + Q_A(x_t)x_t} \leq e^w$$

$$\Rightarrow (a-b)e^w {x_t}^2 + (b - a - ae^w + e^w)x_t + a - e^w \leq 0$$

$$\Rightarrow c = 1 + a^2 - 4b + 2(b-a)(1+a)e^{-w} + (b-a)^2 e^{-2w}$$

$$\Rightarrow x_{1,\pm} = \frac{1}{2e^w} + \frac{a-1}{2(a-b)} \pm \frac{\sqrt{c}}{2(a-b)}$$

$$\Rightarrow x^1_{1,\pm} = -\frac{1}{2e^w} \pm \frac{[2(a-b)(1+a)e^{-w} - 2(b-a)^2 e^{-2w}]}{4(a-b)\sqrt{c}}$$

for $y_t = 0$,

$$\frac{1}{1 - x_t + Q_A(x_t)x_t} \leq e^w$$

$$\Rightarrow (a-b)e^w {x_t}^2 + (1-a)e^w x_t + 1 - e^w \leq 0$$

$$\Rightarrow d = [(1+a)^2 - 4b]e^{2w} + 4(b-a)e^w$$

$$\Rightarrow x_{2,\pm} = \frac{a-1}{2(a-b)} \pm \frac{\sqrt{d}}{2(a-b)e^w}$$

$$\Rightarrow x^1_{2,\pm} = \frac{1}{\sqrt{d}}$$

Since $Q_A(x_t) \geq 0$,$a \geq 0$ and $b \geq 0$.

Firstly, we consider $a, b \leq 1$, that is $Q_A(x_t) \leq 1$.

For $a < b$, we set $w_1 = min_{x \in (0,1)}\{log[\frac{(b-a)x+a}{1-x+(b-a)x^2+ax}]\}$

and $w_2 = max_{x\in(0,1)}\{log[\frac{1}{1-x+(b-a)x^2+ax}]\}$.

We obtain the probability distribution function of W as:

$$f_W(w;\theta,d) = \begin{cases} x^1_{1,+}[\frac{Q(x_{1,+})x_{1,+}}{1-x_{1,+}+Q(x_{1,+})x_{1,+}}]f_X(x_{1,+};\theta), & w_1 \le w < 0; \\ x^1_{2,+}[\frac{1-x_{2,+}}{1-x_{2,+}+Q(x_{2,+})x_{2,+}}]f_X(x_{2,+};\theta) & \\ -x^1_{2,-}[\frac{1-x_{2,-}}{1-x_{2,-}+Q(x_{2,-})x_{2,+}}]f_X(x_{2,-};\theta)I(x_{2,+}<1), & 0 \le w \le w_2, \end{cases} \quad (A11)$$

For $a > b$, we obtain the probability distribution function of W as:

$$f_W(w;\theta,d) = \begin{cases} x^1_{1,+}[\frac{Q(x_{1,+})x_{1,+}}{1-x_{1,+}+Q(x_{1,+})x_{1,+}}]f_X(x_{1,+};\theta) & \\ -x^1_{1,-}[\frac{Q(x_{1,-})x_{1,-}}{1-x_{1,-}+Q(x_{1,-})x_{1,-}}]f_X(x_{1,-};\theta)I(x_{1,-}>0), & w_1 \le w < 0; \\ x^1_{2,+}[\frac{1-x_{2,+}}{1-x_{2,+}+Q(x_{2,+})x_{2,+}}]f_X(x_{2,+};\theta), & 0 \le w \le w_2, \end{cases} \quad (A12)$$

Next, we consider $a, b \ge 1$, that is $Q_A(x_t) \ge 1$.

For $a < b$, we set $w_3 = min_{x\in(0,1)}\{log[\frac{1}{1-x+(b-a)x^2+ax}]\}$

and $w_4 = max_{x\in(0,1)}\{log[\frac{b-a)x+a}{1-x+(b-a)x^2+ax}]\}$.

We obtain the probability distribution function of W as:

$$f_W(w;\theta,d) = \begin{cases} -x^1_{2,-}[\frac{1-x_{2,-}}{1-x_{2,-}+Q(x_{2,-})x_{2,-}}]f_X(x_{2,-};\theta), & w_3 \le w < 0; \\ -x^1_{1,-}[\frac{Q(x_{1,-})x_{1,-}}{1-x_{1,-}+Q(x_{1,-})x_{1,-}}]f_X(x_{1,-};\theta) & \\ +x^1_{1,+}[\frac{Q(x_{1,+})x_{1,+}}{1-x_{1,+}+Q(x_{1,+})x_{1,+}}]f_X(x_{1,+};\theta)I(x_{1,+}>0), & 0 \le w \le w_4; \end{cases} \quad (A13)$$

For $a > b$, we obtain the probability distribution function of W as:

$$f_W(w;\theta,d) = \begin{cases} x^1_{2,+}[\frac{1-x_{2,+}}{1-x_{2,+}+Q(x_{2,+})x_{2,+}}]f_X(x_{2,+};\theta)I(x_{2,+}<1) & \\ -x^1_{2,-}[\frac{1-x_{2,-}}{1-x_{2,-}+Q(x_{2,-})x_{2,-}}]f_X(x_{2,-};\theta), & w_3 \le w < 0; \\ -x^1_{1,-}[\frac{Q(x_{1,-})x_{1,-}}{1-x_{1,-}+Q(x_{1,-})x_{1,-}}]f_X(x_{1,-};\theta), & 0 \le w \le w_4; \end{cases} \quad (A14)$$

Lastly, we consider the case that there exists some $x_0 \in (0,1)$ such that $Q_A(x_0) = (b-a)x_0 + a = 1$, that is $x_0 = (1-a)/(b-a)$.

For $a < b$, we obtain the probability distribution function of W as:

$$f_W(w;\theta,d) = \begin{cases} x^1_{1,+}[\frac{Q(x_{1,+})x_{1,+}}{1-x_{1,+}+Q(x_{1,+})x_{1,+}}]f_X(x_{1,+};\theta)I(x_{1,+} < x_0) \\ -x^1_{2,-}[\frac{1-x_{2,-}}{1-x_{2,-}+Q(x_{2,-})x_{2,-}}]f_X(x_{2,-};\theta)I(x_{2,-} > x_0), \\ \qquad\qquad max(w_1,w_3) \le w < 0; \\ x^1_{2,+}[\frac{1-x_{2,+}}{1-x_{2,+}+Q(x_{2,+})x_{2,+}}]f_X(x_{2,+};\theta)I(x_{2,+} < x_0) \\ -x^1_{2,-}[\frac{1-x_{2,-}}{1-x_{2,-}+Q(x_{2,-})x_{2,-}}]f_X(x_{2,-};\theta)I(x_{2,-} < x_0) \\ -x^1_{1,-}[\frac{Q(x_{1,-})x_{1,-}}{1-x_{1,-}+Q(x_{1,-})x_{1,-}}]f_X(x_{1,-};\theta)I(x_{1,-} > x_0) \\ +x^1_{1,+}[\frac{Q(x_{1,+})x_{1,+}}{1-x_{1,+}+Q(x_{1,+})x_{1,+}}]f_X(x_{1,+};\theta)I(x_{1,+} > x_0), \\ \qquad\qquad 0 \le w \le max(w_2,w_4); \end{cases} \tag{A15}$$

For $a > b$, we obtain the probability distribution function of W as:

$$f_W(w;\theta,d) = \begin{cases} I(x_{1,+} > x_0)[x^1_{1,+}[\frac{Q(x_{1,+})x_{1,+}}{1-x_{1,+}+Q(x_{1,+})x_{1,+}}]f_X(x_{1,+};\theta) \\ -x^1_{1,-}[\frac{Q(x_{1,-})x_{1,-}}{1-x_{1,-}+Q(x_{1,-})x_{1,-}}]f_X(x_{1,-};\theta)I(x_{1,-} > x_0)] \\ +I(x_{2,-} < x_0)[x^1_{2,+}[\frac{1-x_{2,+}}{1-x_{2,+}+Q(x_{2,+})x_{2,+}}]f_X(x_{2,+};\theta)I(x_{2,+} < x_0) \\ -x^1_{2,-}[\frac{1-x_{2,-}}{1-x_{2,-}+Q(x_{2,-})x_{2,-}}]f_X(x_{2,-};\theta)], \\ \qquad\qquad max(w_1,w_3) \le w < 0; \\ I(x_{2,+} > x_0)x^1_{2,+}[\frac{1-x_{2,+}}{1-x_{2,+}+Q(x_{2,+})x_{2,+}}]f_X(x_{2,+};\theta) \\ -I(x_{1,-} < x_0)x^1_{1,-}[\frac{Q(x_{1,-})x_{1,-}}{1-x_{1,-}+Q(x_{1,-})x_{1,-}}]f_X(x_{1,-};\theta), \\ \qquad\qquad 0 \le w \le max(w_2,w_4); \end{cases} \tag{A16}$$

APPENDIX D. INVESTIGATION OF BANDWIDTH PARAMETERS

In the discussion of using kernel-based matching estimators, under the contemplation of giving higher weights on patients close in terms of the mortality risk x_t whilst lower weights on more distant observations, the kernel function $K(\cdot)$ can be chosen to be a symmetric, nonnegative, unimodal kernel, typically the Gaussian with mean of 0 and variance of 1. The Gaussian kernel function can be streamlined to obtain robustness to extreme outliers by limiting the support of the kernel, such as setting it to 0 for distances greater than 2. Alternatively, the kernel function can be chosen to be the cosine or Epanechnikov functions. Silverman (1981) showed that apart from the theoretical advantages of using the Gaussian kernel function, such as the inheritance of continuity and differentiability properties for the estimation, it also has strong computational advantages, such as not involving any nonlinear optimization procedures.

Consequently, Silverman (1986) stated a "rule of thumb" for the selection of the optimal bandwidth for using the Gaussian kernel function is $h = 0.9n^{-1/5}$ $\min\{s, IQR/1.34\}$ where IQR is the sample interquartile range and s is the sample standard deviation. This is suggested because it will, to a high degree of accuracy, minimize the integrated mean square error, as shown in Deheuvels (1977). For many situations, this will be an adequate choice of the bandwidth but for distributions that have relatively large variance with a small range of preliminary observations, Chen and Kelton (2006) suggested a minor adjustment, in which $h = 0.9n^{-1/5}\min\{s, IQR/2.68\}$. The bandwidth can also be determined using plug-in methods by Park and Marron (1990), and Sheather and Jones (1991).

Though the above suggested bandwidths are initially introduced for the estimation of Kernel densities, the bandwidths describe the width of the convolution kernel used. To a layman, the bandwidth will be an indication of the number of observations that is to be considered in the neighborhood so as to obtain "an adequate estimator". As such, we will like to test if the above suggested bandwidths are applicable in our context.

Using various values of the bandwidth h for the Gaussian kernel function, examples of the estimates $\hat{p}(x_t; h)$ for a sample data set of size $n = 1000$, one under $Q = 2$ and the other under $Q = 0.5$ are displayed in Figures A1 and A2 respectively. The mortality risk x_t is simulated from beta distribution which is parameterized by shape parameters $\alpha = 1$ and $\beta = 3$, while the outcome y_t is simulated from Bernoulli distribution parameterized by $p(x_t) = Qx_t/(1-x_t+Qx_t)$. The basis for determining the parameters and various aspects of the examples as above is to enable us to simulate examples with distributional characteristics of the mortality risk which mimics that of a real data set we have discussed.

Broadly speaking, from Figures Figures A1(a)-(c) and Figures A2(a)-(c), a kernel estimator is likely to under- and oversmooth $p(x_t)$ by using various values of the bandwidth h. The choice of the bandwidth is known to generally involve a trade-off between variance and bias of the estimator. If a small bandwidth is used, such as in our examples where $h = 0.01$ (in Figures A1(a) and A2(a)), we tend to be able to capture local characteristics of $p(x_t)$ but we will not be able to obtain global characteristics of $p(x_t)$. This translates to eliminating the bias but the resulting variance will be large. By using a large bandwidth, such as in our

example where $h = 0.2$ (in Figures A1(b) and A2(b)), the global characteristics of $p(x_t)$ can be obtained but we will lose information of its local characteristics. Consequently, the variance will be reduced but the bias will also be increased. As such, we need to evaluate the quality of our estimates and address the above trade-off. For each value of bandwidth h considered in Figures A1(a)-(c) and Figures A2(a)-(c), a smoother estimate of the odds ratio function using the MSE criterion is obtained and this is used to compute $\hat{p}(x_t)$. We observe that the estimates $\hat{p}(x_t)$ are closest to the true $p(x_t)$ through the use of the adjusted bandwidth $h = 0.9n^{-1/5}\min\{s, IQR/2.68\}$ suggested by Chen and Kelton (2006). This shows the applicability of this bandwidth in our context.

The plots of odds ratio against the mortality risk x_t in Figures A1(d) and A2(d) also suggest that the odds ratios are indeed constants. The high values of odds ratio observed for small values of x_t in Figure A2(d) is inherent, due to large values of $x_t/(1 - x_t)$. As a result, a very small increase in the estimate $\hat{p}(x_t; h)$ tends to result in a much higher value of odds ratio.

We conduct a simulation study to further show the applicability of the bandwidths suggested by Silverman (1986), and Chen and Kelton (2006) in our context. For $Q = 2$ and 0.5, the simulation was replicated 1000 times, resulting in 1000 data sets with each data set comprising $n = 1000$ subjects. The mortality risk x_t for each subject was drawn from a beta distribution with shape parameters $\alpha = 1$ and $\beta = 2, 2.5, 3, 4$ or 5, while the corresponding discrete outcome was generated from a Bernoulli distribution with $p(x_t) = Qx_t/(1 - x_t + Qx_t)$.

Table A1 contains the results of the simulation study. The values of the

optimal bandwidth h are obtained by finding the value of h that estimates $p(x_t; h)$ which upon the minimization of (3.6), gives an estimate of $\hat{Q}$ closest to the true Q. This is done through the use of a tedious grid search. We observe that as the distribution becomes less skewed to the right (as the shape parameter β increases), there are more subjects with lower mortality risk and less subjects with higher mortality risk. As a result, the optimal bandwidth to attain the smallest MSE decreases. This is readily interpreted because most of the subjects have lower mortality risk, thus in order to better estimate $p(x_t)$ in that region, the required bandwidth need not be comparatively large. Moreover, we observe that through the use of the adjusted bandwidth h_2 suggested by Chen and Kelton (2006), it emanates $\hat{Q}$ at a performance similar to that when using the optimal bandwidth h. These results support the rationale of using $h = 0.9n^{-1/5}\min\{s, IQR/2.68\}$ in our context.

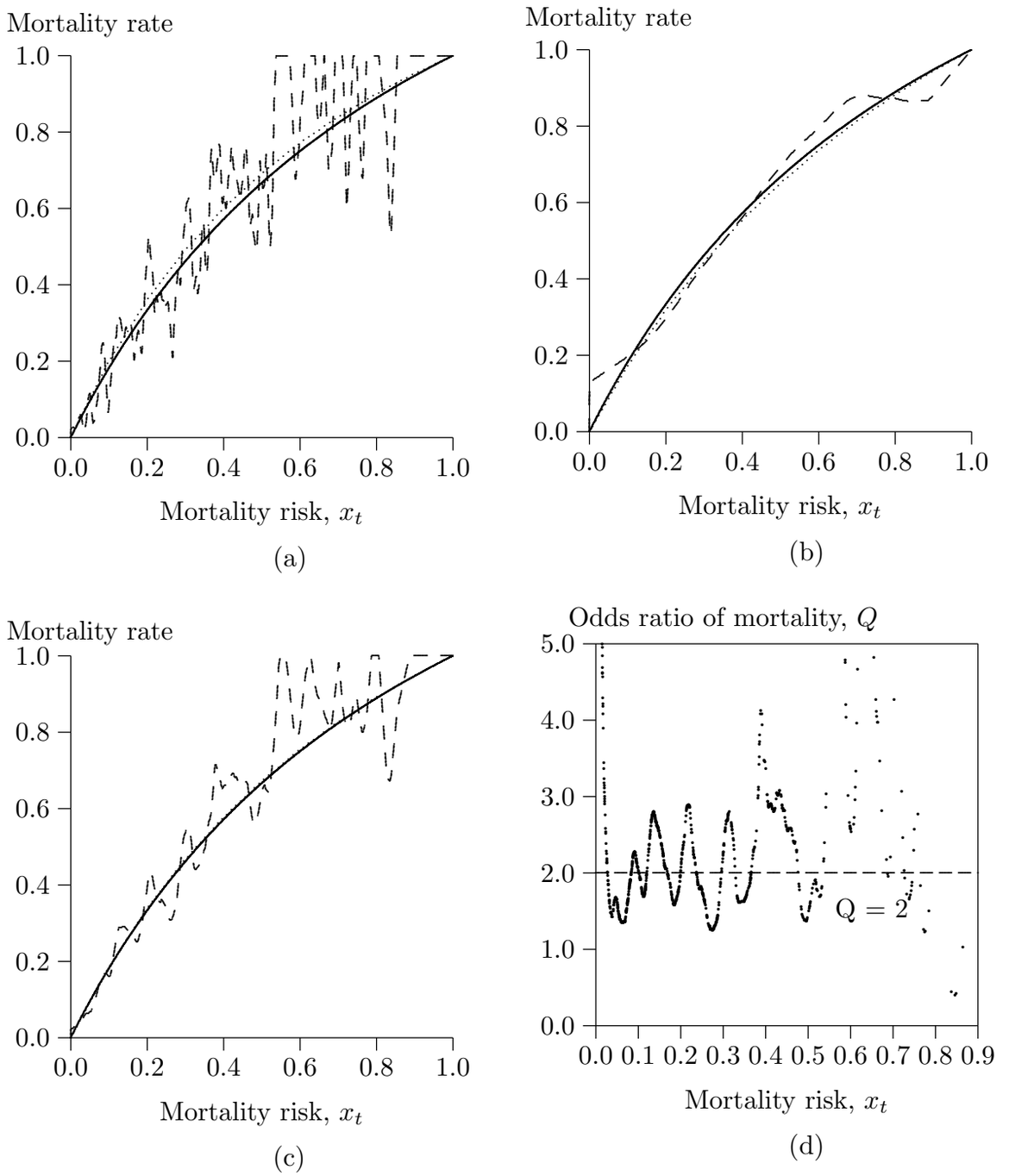

Figure A1. Unsmoothed Kernel estimate $\hat{p}(x_t; h)$(equation (3.4), represented by dashed line), smoothed MSE estimate $\hat{p}(x_t)$ (using equation (3.6), represented by dotted line) of mortality rate with simulated data of size $n = 1000$ for (a) $h = 0.01$, (b) $h = 0.2$ and (c) $h = 0.9\ n^{-1/5} \min\{s,\ IQR/2.68\}$ under $Q = 2$. The true mortality rate $p(x_t)$ is represented by the solid line. Note that the dotted line and the solid line is almost perfectly matched. For $\hat{p}(x_t; h)$ obtained using (c), the plot of odds ratio of mortality Q against mortality risk x_t is shown in (d).

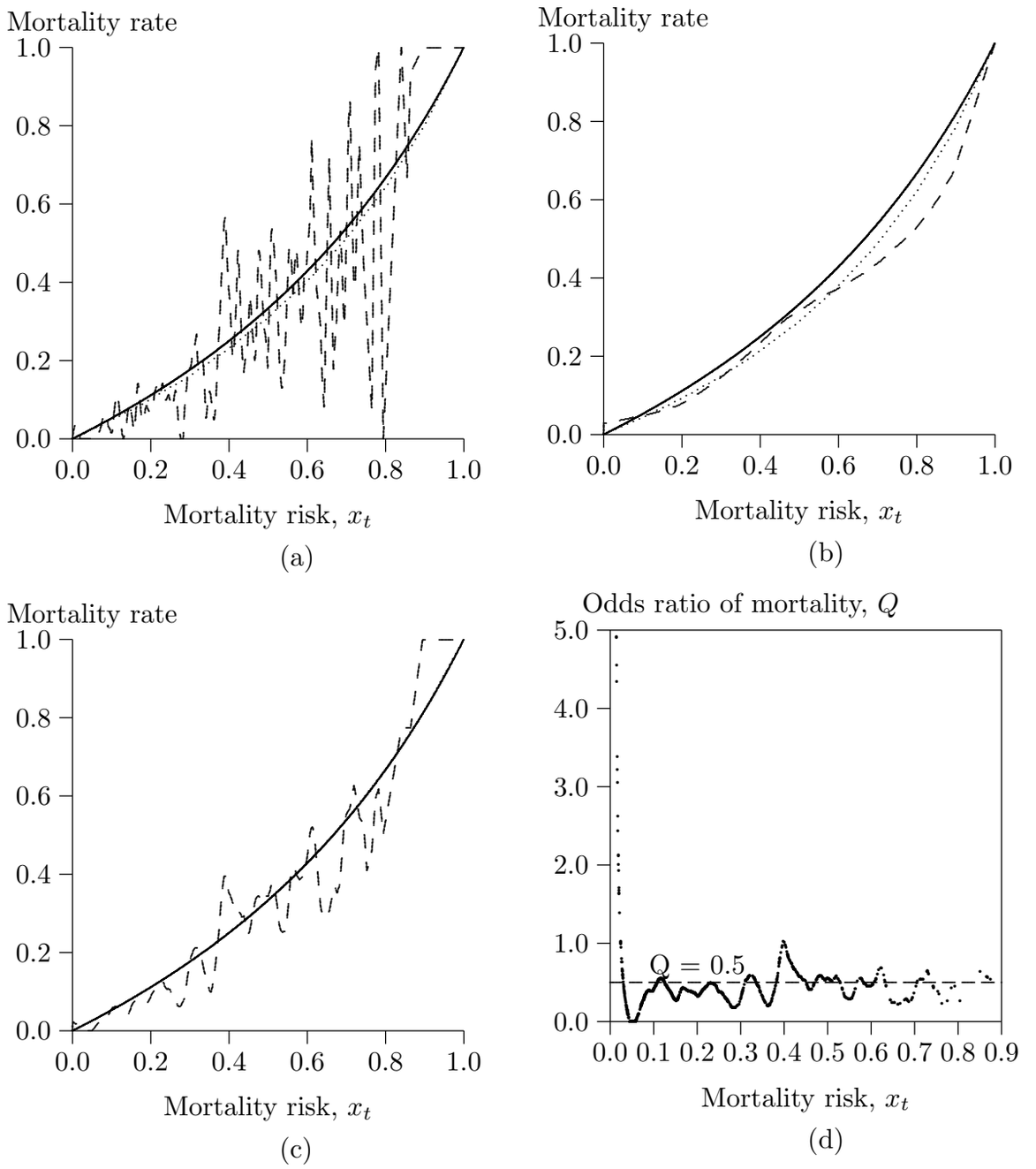

Figure A2. Unsmoothed Kernel estimate $\hat{p}(x_t; h)$(equation (3.4), represented by dashed line), smoothed MSE estimate $\hat{p}(x_t)$ (using equation (3.6), represented by dotted line) of mortality rate with simulated data of size $n = 1000$ for (a) $h = 0.01$, (b) $h = 0.2$ and (c) $h = 0.9\ n^{-1/5} \min\{s,\ IQR/2.68\}$ under $Q = 0.5$. The true mortality rate $p(x_t)$ is represented by the solid line. Note that the dotted line and the solid line is almost perfectly matched. For $\hat{p}(x_t; h)$ obtained using (c), the plot of odds ratio of mortality Q against mortality risk x_t is shown in (d).

Table A1. Analysis of $\hat{Q}$ and its corresponding standard errors using optimal (h), Silverman (1986)'s (h_1) and, Chen and Kelton (2006)'s (h_2) bandwidths.

True Q	Distribution	Optimal h	$\hat{Q}$ (SE^*)	h_1 (SE^*)	$\hat{Q}_1$ (SE^*)	h_2 (SE^*)	$\hat{Q}_2$ (SE^*)
2.0	beta(1,2)	0.0476	2.0000	0.0533	1.9978	0.0307	2.0056
			(0.0048)	(<0.0001)	(0.0048)	(<0.0001)	(0.0048)
	beta(1,2.5)	0.0264	2.0000	0.0483	1.9902	0.0267	2.0000
			(0.0042)	(<0.0001)	(0.0041)	(<0.0001)	(0.0042)
	beta(1,3)	0.0207	1.9847	0.0439	1.9684	0.0233	1.9783
			(0.0048)	(<0.0001)	(0.0047)	(<0.0001)	(0.0047)
	beta(1,4)	0.0159	1.9894	0.0365	1.9722	0.0186	1.9840
			(0.0052)	(<0.0001)	(0.0051)	(<0.0001)	(0.0051)
	beta(1,5)	0.0107	1.9928	0.0309	1.9778	0.0156	1.9882
			(0.0061)	(<0.0001)	(0.0058)	(<0.0001)	(0.0058)
0.5	beta(1,2)	0.0501	0.4972	0.0533	0.4942	0.0307	0.5059
			(0.0014)	(<0.0001)	(0.0014)	(<0.0001)	(0.0014)
	beta(1,2.5)	0.0490	0.5000	0.0483	0.5045	0.0267	0.5065
			(0.0014)	(<0.0001)	(0.0014)	(<0.0001)	(0.0014)
	beta(1,3)	0.0381	0.5000	0.0439	0.4994	0.0233	0.5011
			(0.0016)	(<0.0001)	(0.0016)	(<0.0001)	(0.0017)
	beta(1,4)	0.0298	0.5000	0.0365	0.4992	0.0186	0.5012
			(0.0017)	(<0.0001)	(0.0017)	(<0.0001)	(0.0018)
	beta(1,5)	0.0153	0.5000	0.0309	0.4974	0.0156	0.4989
			(0.0020)	(<0.0001)	(0.0020)	(<0.0001)	(0.0020)

SE^*: Standard error of estimate