Development of Statistical Methods for the Surveillance and Monitoring of Adverse Events which adjust for Differing Patient and Surgical Risks

Ronald Albert Webster
Bachelor of Applied Science (Mathematics), QUT
Bachelor of Applied Science (Honours), QUT

A thesis submitted in partial fulfilment for the degree of

Doctor of Philosophy
February 2008

Principal Supervisor: Professor Anthony N Pettitt
Associate Supervisor: Dr Helen L Johnson
Industry Supervisor: Dr Anthony P Morton

Queensland University of Technology
School of Mathematical Sciences
Faculty of Science
Brisbane, Queensland, 4000, AUSTRALIA
Keywords

Adverse outcomes, ARL, control chart, Markov chain property, medical monitoring, minimum effect CUSUM, meta-factors, patient level uncertainty, parameter estimation uncertainty, risk adjustment, risk adjusted CUSUM, report cards, risk score, risk model performance.
Abstract

The research in this thesis has been undertaken to develop statistical tools for monitoring adverse events in hospitals that adjust for varying patient risk. The studies involved a detailed literature review of risk adjustment scores for patient mortality following cardiac surgery, comparison of institutional performance, the performance of risk adjusted CUSUM schemes for varying risk profiles of the populations being monitored, the effects of uncertainty in the estimates of expected probabilities of mortality on performance of risk adjusted CUSUM schemes, and the instability of the estimated average run lengths of risk adjusted CUSUM schemes found using the Markov chain approach.

The literature review of cardiac surgical risk found that the number of risk factors in a risk model and its discriminating ability were independent, the risk factors could be classified into their “dimensions of risk”, and a risk score could not be generalized to populations remote from its developmental database if accurate predictions of patients’ probabilities of mortality were required. The conclusions were that an institution could use an “off the shelf” risk score, provided it was recalibrated, or it could construct a customized risk score with risk factors that provide at least one measure for each dimension of risk.

The use of report cards to publish adverse outcomes as a tool for quality improvement has been criticized in the medical literature. An analysis of the report cards for cardiac surgery in New York State showed that the institutions’ outcome rates appeared overdispersed compared to the model used to construct confidence intervals, and the uncertainty associated with the estimation of institutions’ out-
come rates could be mitigated with trend analysis. A second analysis of the mortality of patients admitted to coronary care units demonstrated the use of notched box plots, fixed and random effect models, and risk adjusted CUSUM schemes as tools to identify outlying hospitals. An important finding from the literature review was that the primary reason for publication of outcomes is to ensure that health care institutions are accountable for the services they provide.

A detailed review of the risk adjusted CUSUM scheme was undertaken and the use of average run lengths (ARLs) to assess the scheme, as the risk profile of the population being monitored changes, was justified. The ARLs for in-control and out-of-control processes were found to increase markedly as the average outcome rate of the patient population decreased towards zero. A modification of the risk adjusted CUSUM scheme, where the step size for in-control to out-of-control outcome probabilities were constrained to no less than 0.05, was proposed. The ARLs of this “minimum effect” CUSUM scheme were found to be stable.

The previous assessment of the risk adjusted CUSUM scheme assumed that the predicted probability of a patient’s mortality is known. A study of its performance, where the estimates of the expected probability of patient mortality were uncertain, showed that uncertainty at the patient level did not affect the performance of the CUSUM schemes, provided that the risk score was well calibrated. Uncertainty in the calibration of the risk model appeared to cause considerable variation in the ARL performance measures.

The ARLs of the risk adjusted CUSUM schemes were approximated using simulation because the approximation method using the Markov chain property of CUSUMs, as proposed by Steiner et al. (2000), gave unstable results. The cause of the instability was the method of computing the Markov chain transition probabilities, where probability is concentrated at the midpoint of its Markov state. If probability was assumed to be uniformly distributed over each Markov state, the ARLs were stabilized, provided that the scores for the patients’ risk of adverse outcomes were discrete and finite.
### Contents

1 Overview of the Research and Investigations ................................. 1

2 Literature Review .................................................................... 7

   2.1 Standard Control Charts ................................................. 7
       2.1.1 For Numerical Data .............................................. 8
       2.1.2 For Count Data .................................................. 9
       2.1.3 For Proportion Data ............................................. 11

   2.2 Monitoring Adverse Events ............................................. 12
       2.2.1 Rare Events ....................................................... 12
       2.2.2 Risk Adjusted Charts ......................................... 14
       2.2.3 Comparing Institutions ....................................... 17
       2.2.4 Performance Evaluation ...................................... 19

   2.3 Risk Scores for Trauma Victims ....................................... 20
       2.3.1 Triage Scores .................................................... 24
       2.3.2 Quality Improvement .......................................... 29

   2.4 Conclusion .................................................................. 34

3 Meta-Factors for Cardiac Surgical Risk Models ............................ 37

   3.1 Introduction .................................................................. 38
   3.2 Three Risk Models ...................................................... 43
       3.2.1 Simple Classification .......................................... 43
       3.2.2 Parsonnet Score ............................................... 44
5.1.2 Schemes for Mortality Outcomes ........................................ 112
5.2 Review of Assessment Methods ........................................... 116
5.3 Assessing Performance Using the ARL .............................. 121
  5.3.1 Population Distributions .............................................. 122
  5.3.2 Definition of the Risk Adjusted CUSUM .......................... 126
  5.3.3 Performance of the Risk Adjusted CUSUM ...................... 127
5.4 Properties of Risk Adjusted CUSUMs ............................... 134
  5.4.1 Mean & Variance of Weights ........................................ 135
  5.4.2 Effect Size ............................................................ 139
5.5 Risk Adjusted & Minimum Effect CUSUMs .......................... 142
5.6 Industrial & Risk Adjusted CUSUMs ................................ 153
5.7 Conclusion & Recommendation ....................................... 156

6 Performance of Risk Adjusted CUSUM Schemes: Expected Outcome Probability Uncertain 159
  6.1 Introduction ............................................................. 159
  6.2 Uncertainty at Patient Level ......................................... 161
    6.2.1 Beta Distribution—$p_T$ Unbiased .............................. 164
    6.2.2 Gamma Distribution—Odds Unbiased ......................... 165
    6.2.3 Gamma Distribution—$p_T$ Unbiased ........................... 166
    6.2.4 Normal Distribution—Logit Unbiased ......................... 167
    6.2.5 Normal Distribution—$p_T$ Unbiased ........................... 167
  6.3 ARLs with Patient Level Uncertainty .............................. 168
  6.4 Effect of Patient Level Uncertainty .............................. 193
  6.5 Uncertainty in Model Parameter Estimates ...................... 196
    6.5.1 Estimated Probability of Outcome .............................. 197
    6.5.2 ARL$_0$s of Risk Adjusted CUSUMs ......................... 201
    6.5.3 Effect on Risk Adjusted CUSUM Alarms ...................... 204
  6.6 Conclusion ............................................................. 210
  6.7 Addendum .............................................................. 212
List of Figures

3.1 Each of the AROC values given for the risk models in Table 3.4 plotted against the number of risk factors used to construct the risk score ................................................. 59

4.1 Trends of the RAMRs for CABG surgery and associated 95% confidence intervals for two hospitals in the New York State Health Department System .................................. 91

4.2 Box plots showing the patient distributions in some Coronary Care Units in north-east England stratified according to the patients’ expected probability of mortality .......................... 97

4.3 Upward and Downward CUSUM Charts monitoring post-admission mortality in the coronary care units of four hospitals in north-east England .................................................. 102

5.1 Example of an industrial CUSUM scheme ........................................... 111

5.2 Example of a risk adjusted CUSUM scheme ......................................... 115

5.3 Distributions of cardiac surgical populations classified by risk of postoperative mortality given by the Parsonnet score ............................... 123

5.4 Distributions of cardiac surgical populations classified by risk of postoperative mortality given by the EuroSCORE ........................................... 125

5.5 Examples showing the relationship of risk adjusted CUSUM weights with expected probabilities of mortality ......................................................... 128
5.6

Plots illustrating the increase in the ARL0 s of risk adjusted CUSUM
schemes with decreasing average expected probability of mortality . 130

5.7

Plots illustrating the increase in the ARL1 s of risk adjusted CUSUM
schemes with decreasing average expected probability of mortality . 132

5.8

Examples to illustrate the changes in the mean E(W ) and variance
Var(W ) of the risk adjusted CUSUM weights as the expected probability of mortality p varies . . . . . . . . . . . . . . . . . . . . . . . 137

5.9

Plots to illustrate the variation in the absolute effect size, to which
a risk adjusted CUSUM is tuned, with changes in the expected
probability of mortality . . . . . . . . . . . . . . . . . . . . . . . . . 141

5.10 Plots illustrating the differences between the absolute effect sizes of
minimum effect and risk adjusted CUSUM schemes . . . . . . . . . 143
5.11 Plots to show the differences between the ARL1 s of minimum effect
and risk adjusted CUSUMs, for the ARL0 s of both schemes fixed at
5,000 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 149
5.12 Decision thresholds for ARL0 s = 5,000 of minimum effect and risk
adjusted CUSUM schemes plotted against the mean expected probabilities of mortality of the populations being monitored . . . . . . 151
5.13 Comparison of a risk adjusted and a minimum effect CUSUM scheme
monitoring mortality in the ICU, Princess Alexandra Hospital . . . 152
5.14 Plot illustrating the contrasting behaviour of the ARL0 s of CUSUM
schemes with scaled weights and schemes with unscaled weights . . 155
6.1

ARL0 s of minimum effect CUSUM schemes where deaths were simulated from “true” probabilities of mortality, p T , drawn from a
Beta(α, β) distribution such that E(p T ) = pm

6.2

. . . . . . . . . . . . 171

ARL0 s of risk adjusted CUSUM schemes where deaths were simulated from “true” probabilities of mortality, p T , drawn from a
Beta(α, β) distribution such that E(p T ) = pm

6.3

. . . . . . . . . . . . 172

ARL0 s of minimum effect CUSUM schemes where deaths were simulated from “true” odds of mortality, Odds T , drawn from a Gamma(α, β)

distribution such that E Odds T = Oddsm . . . . . . . . . . . . . . 173
x


6.4 ARLs of risk adjusted CUSUM schemes where deaths were simulated from “true” odds of mortality, Odds\(\text{T}\), drawn from a Gamma\((\alpha, \beta)\) distribution such that \(E(\text{Odds}_T) = \text{Odds}_m\) . . . . . . . . . . . . . . . . . . . . . . . . . 174

6.5 ARLs of minimum effect CUSUM schemes where deaths were simulated from “true” odds of mortality, Odds\(\text{T}\), drawn from a Gamma\((\alpha, \beta)\) distribution such that \(E(p_T) \approx p_m\) . . . . . . . . . . . . . . . . . . . . . . . . . 176

6.6 ARLs of risk adjusted CUSUM schemes where deaths were simulated from “true” odds of mortality, Odds\(\text{T}\), drawn from a Gamma\((\alpha, \beta)\) distribution such that \(E(p_T) \approx p_m\) . . . . . . . . . . . . . . . . . . . . . . . . . 177

6.7 ARLs of minimum effect CUSUM schemes where deaths were simulated from logit\(p_T\)s drawn from an \(N(\text{logit}(p_m), \sigma^2)\) distribution such that \(E(\text{logit}(p_T)) = \text{logit}(p_m)\) . . . . . . . . . . . . . . . . . . . . . . . . . 179

6.8 ARLs of risk adjusted CUSUM schemes where deaths were simulated from logit\(p_T\)s drawn from an \(N(\text{logit}(p_m), \sigma^2)\) distribution such that \(E(\text{logit}(p_T)) = \text{logit}(p_m)\) . . . . . . . . . . . . . . . . . . . . . . . . . 180

6.9 ARLs of minimum effect CUSUM schemes where deaths were simulated from logit\(p_T\)s drawn from an \(N(\text{logit}(p_m), \sigma^2)\) distribution such that \(E(p_T) \approx p_m\) . . . . . . . . . . . . . . . . . . . . . . . . . 182

6.10 ARLs of risk adjusted CUSUM schemes where deaths were simulated from logit\(p_T\)s drawn from an \(N(\text{logit}(p_m), \sigma^2)\) distribution such that \(E(p_T) \approx p_m\) . . . . . . . . . . . . . . . . . . . . . . . . . 183

6.11 ARLs of risk adjusted and minimum effect CUSUM schemes monitoring deaths in a patient population with mean expected probability of mortality \(\bar{p} = 0.020\) . . . . . . . . . . . . . . . . . . . . . . . . . 190

6.12 ARLs of risk adjusted and minimum effect CUSUM schemes monitoring deaths in a patient population with mean expected probability of mortality \(\bar{p} = 0.051\) . . . . . . . . . . . . . . . . . . . . . . . . . 191

6.13 ARLs of risk adjusted and minimum effect CUSUM schemes monitoring deaths in a patient population with mean expected probability of mortality \(\bar{p} = 0.160\) . . . . . . . . . . . . . . . . . . . . . . . . . 192

6.14 Distributions of ARLs estimated with uncertain \((\hat{\alpha}, \hat{\beta})^T\) plotted against the size of training data sets used to estimate \((\hat{\alpha}, \hat{\beta})^T\) . . . . . . . . . . . . . . . . . . . . . . . . . 203
6.15 Examples to illustrate the influence of calibration of the APACHE III score on the path of upward and downward CUSUMs monitoring outcomes in an ICU .................................................. 206

6.16 Additional examples to illustrate the influence of calibration of the APACHE III score on the path of upward and downward CUSUMs monitoring outcomes in an ICU ........................................ 207

7.1 Study using artificial risk adjustment. ARLs of risk adjusted CUSUMs found using the Markov chain approach versus the number of discretization of the decision interval ........................................ 223

7.2 Two cardiac surgical populations. One classified by Parsonnet score and the other by EuroSCORE .......................................................... 225

7.3 Study using cardiac surgical risk scores. ARLs of risk adjusted CUSUMs found using the Markov chain approach versus number of discretizations of the decision interval ........................... 227
List of Tables

3.1 Risk factors and weights for Parsonnet score .......................... 45
3.2 Risk factors, risk factor descriptions, risk factor weights, and parameter estimates for the EuroSCORE .......................... 48
3.3 Proposed meta-factor categories for cardiac surgical outcomes ... 54
3.4 The number of contributing data sources, area under the receiver operating characteristic curve, the number of risk factors by meta-factor class, and total number of risk factors for 27 cardiac surgical risk scores .......................................................... 55
3.5 Assessment measures for the Parsonnet score. Number of data sources, area under the receiver operating characteristic curve, mean predicted probability of mortality, observed mortality rate, and standardized mortality rate .......................................................... 62
3.6 Assessment measures for the EuroSCORE. Number of data sources, area under the receiver operating characteristic curve, mean predicted probability of mortality, observed mortality rate, and standardized mortality rate .......................................................... 64
4.1 Number and percentage of New York State hospitals with mortality rates following CABG surgery which differ significantly from the state average .......................................................... 89
4.2 Standardized mortality rates with credible intervals for fixed and random effects for acute myocardial infarction patients admitted to 11 Yorkshire hospitals .......................................................... 96
5.1 Results of simulation studies to approximate $\text{ARL}_0$s of a risk adjusted CUSUM scheme .......................................................... 129

5.2 Results of simulation studies to approximate $\text{ARL}_1$s of a risk adjusted CUSUM scheme .......................................................... 131

5.3 Minimum means $\text{E}(W)$ and maximum standard deviations $\text{SD}(W)$ and variances $\text{Var}(W)$ of the weight of the risk adjusted CUSUM scheme as the odds ratio, $R_A$, diverges from one ............................................ 138

5.4 Results of simulation studies to approximate $\text{ARL}_0$s of a risk adjusted CUSUM scheme with a minimum effect size of 5% ......................................................... 145

5.5 Results of simulation studies to approximate $\text{ARL}_1$s of a risk adjusted CUSUM scheme with a minimum effect size of 5% ......................................................... 146

5.6 Results of simulation studies to approximated the $\text{ARL}_1$s of a risk adjusted CUSUM scheme, where the out-of control mortality rate is as for the minimum effect CUSUM scheme ........................................... 150

5.7 Results of an exploratory simulation study to compare and contrast approximated $\text{ARL}_0$s of scaled and unscaled Bernoulli CUSUM schemes ............................................................. 154

6.1 Ratios of $\text{ARL}_0$s approximated by the Gamma model given in Section 6.2.2 and $\text{ARL}_0$s found using the approximation process described in Chapter 5 ......................................................... 175

6.2 Ratios of $\text{ARL}_0$s approximated using the Gamma correction model given in Section 6.2.3 and $\text{ARL}_0$s found using the approximation process described in Chapter 5 ......................................................... 178

6.3 Ratios of $\text{ARL}_0$s approximated using the Normal model given in Section 6.2.4 and $\text{ARL}_0$s found using the approximation process described in Chapter 5 ......................................................... 181

6.4 Ratios of $\text{ARL}_0$s approximated using the Normal correction model given in Section 6.2.5 and $\text{ARL}_0$s found using the approximation process described in Chapter 5 ......................................................... 184
6.5 The means and standard deviations of the expected probabilities of mortality, found using Taylor series approximations, after recalibrating the Parsonnet score and EuroSCORE by fitting Equation (6.29) to data sets of size $n$.

6.6 Parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, for recalibration of the APACHE III score. The estimates were derived by fitting training data sets of size $n$ to Equation (6.29).

7.1 Maximum relative differences between ARL approximations found using simulation and those found using the Markov chain approach with transition probabilities calculated using rounding.
# List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACSQHC</td>
<td>Australian Council for Safety and Quality in Health Care</td>
</tr>
<tr>
<td>AFT</td>
<td>Accelerated Failure Time</td>
</tr>
<tr>
<td>AIC</td>
<td>Akaike Information Criterion</td>
</tr>
<tr>
<td>AIS</td>
<td>Abbreviated Injury Score</td>
</tr>
<tr>
<td>AMI</td>
<td>Acute Myocardial Infarct</td>
</tr>
<tr>
<td>APACHE</td>
<td>Acute Physiology And Chronic Health Evaluation</td>
</tr>
<tr>
<td>ARC</td>
<td>Australian Research Council</td>
</tr>
<tr>
<td>ARL</td>
<td>Average Run Length</td>
</tr>
<tr>
<td>$ARL_0$</td>
<td>Average Run Length when the process is in-control</td>
</tr>
<tr>
<td>$ARL_1$</td>
<td>Average Run Length when the process is out-of-control</td>
</tr>
<tr>
<td>AROC</td>
<td>Area under the Receiver Operating Characteristic</td>
</tr>
<tr>
<td>ASN</td>
<td>Average Sample Number</td>
</tr>
<tr>
<td>AVR</td>
<td>Aortic Valve Replacement</td>
</tr>
<tr>
<td>BP</td>
<td>Blood Pressure</td>
</tr>
<tr>
<td>CABG</td>
<td>Coronary Artery Bypass Graft</td>
</tr>
<tr>
<td>CAD</td>
<td>Coronary Artery Disease</td>
</tr>
<tr>
<td>CCU</td>
<td>Coronary Care Unit</td>
</tr>
<tr>
<td>CHF</td>
<td>Congestive Heart Failure</td>
</tr>
<tr>
<td>COPD</td>
<td>Chronic Obstructive Pulmonary Disease</td>
</tr>
<tr>
<td>CRAM</td>
<td>Cumulative Risk Adjusted Mortality</td>
</tr>
<tr>
<td>CSRS</td>
<td>Cardiac Surgery Reporting System</td>
</tr>
<tr>
<td>CUSUM</td>
<td>Cumulative Sum</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>DDD</td>
<td>Defined Daily Dose</td>
</tr>
<tr>
<td>DRG</td>
<td>Diagnosis Related Group</td>
</tr>
<tr>
<td>EMMACE</td>
<td>Evaluation of Methods and Management of Acute Coronary Events</td>
</tr>
<tr>
<td>EMR</td>
<td>Expected Mortality Rate</td>
</tr>
<tr>
<td>ESBL</td>
<td>Extended Spectrum Beta Lactamase</td>
</tr>
<tr>
<td>EuroSCORE</td>
<td>European System for Cardiac Operative Risk Evaluation</td>
</tr>
<tr>
<td>EWMA</td>
<td>Exponentially Weighted Moving Average</td>
</tr>
<tr>
<td>HDL</td>
<td>High Density Lipoprotein</td>
</tr>
<tr>
<td>IABP</td>
<td>Intra-Aortic Balloon Pump</td>
</tr>
<tr>
<td>ICU</td>
<td>Intensive Care Unit</td>
</tr>
<tr>
<td>IMA</td>
<td>Internal Mammary Artery</td>
</tr>
<tr>
<td>LOS</td>
<td>Length Of Stay</td>
</tr>
<tr>
<td>LV</td>
<td>Left Ventricular</td>
</tr>
<tr>
<td>LVEF</td>
<td>Left Ventricular Ejection Fraction</td>
</tr>
<tr>
<td>MCI</td>
<td>Multiple Casualty Incidents</td>
</tr>
<tr>
<td>MINAP</td>
<td>Myocardial Infarction National Audit Project</td>
</tr>
<tr>
<td>MRO</td>
<td>Multiply Resistant Organism</td>
</tr>
<tr>
<td>MRSA</td>
<td>Methicillin Resistant <em>Staphylococcus aureus</em></td>
</tr>
<tr>
<td>MVR</td>
<td>Mitral Valve Replacement</td>
</tr>
<tr>
<td>NHS</td>
<td>National Health Service</td>
</tr>
<tr>
<td>NNIS</td>
<td>National Nosocomial Infection Surveillance</td>
</tr>
<tr>
<td>NYHA</td>
<td>New York Heart Association</td>
</tr>
<tr>
<td>OMR</td>
<td>Observed Mortality Rate</td>
</tr>
<tr>
<td>OR</td>
<td>Operating Room</td>
</tr>
<tr>
<td>PA</td>
<td>Pulmonary Artery (pressure)</td>
</tr>
<tr>
<td>PICU</td>
<td>Paediatric Intensive Care Unit</td>
</tr>
<tr>
<td>PIM2</td>
<td>Paediatric Index of Mortality 2</td>
</tr>
<tr>
<td>PTCA</td>
<td>Percutaneous Transluminal Coronary Angioplasty</td>
</tr>
<tr>
<td>RAMR</td>
<td>Risk Adjusted Mortality Rate</td>
</tr>
<tr>
<td>RAST</td>
<td>Risk Adjusted Survival Time</td>
</tr>
<tr>
<td>RIDIT</td>
<td>Relative to an Identified Distribution</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>RSPRT</td>
<td>Resetting Sequential Probability Ratio Test</td>
</tr>
<tr>
<td>SPC</td>
<td>Statistical Process Control</td>
</tr>
<tr>
<td>SPRT</td>
<td>Sequential Probability Ratio Test</td>
</tr>
<tr>
<td>VLAD</td>
<td>Variable Life Adjusted Display</td>
</tr>
</tbody>
</table>
Statement of Original Ownership

The work contained in this thesis has not been submitted previously for a degree or diploma at any other higher education institution. To the best of my knowledge and belief, the thesis contains no material published previously or written by another person, except where due reference is made.

Signed: __________________________

Date: __________________________
Acknowledgments

This thesis is dedicated to my late mother, Betty Patricia Webster, who passed away while I was undertaking the research for this degree. I have tried to emulate her modesty and quiet determination to do her best.

There were many who supported me during the course of study and writing of this thesis. First, thanks to my sisters and brother who encouraged me during the hard times and to my father, now in the sunset of his life, who ensured I had a good education and a grasp of reading and arithmetic.

Second, I am grateful to my supervisory team and members of the ARC Hospital Linkage team for their support. My principal supervisor, Professor Tony Pettitt, offered me the chance to do this study and advised on the use of statistical methodologies. My industry supervisor, Dr Tony Morton, ensured I understood and used the Deming philosophy of quality process control for this study. Chapter 3 of this thesis was written in collaboration with Dr David Cook. He suggested the meta-factor categories and classified the risk factors. My associate supervisor, Dr Helen Johnson, provided support after Tony Pettit went on leave to Lancaster University. My thanks to all at the School of Mathematical Sciences, QUT, who helped. I shall miss the camaraderie of the PhD room.

Finally, I wish to thank the Australian Research Council for providing the funding for my Australian Postgraduate Industry Award (APIA) scholarship. I would have been unable to undertake the study without that financial support.
Chapter 1

Overview of the Research and Investigations

The research for this thesis was undertaken as part of a project for an Australian Research Council (ARC) Linkage scheme, called *Statistical and mathematical modelling to improve outcomes in health care*, involving the School of Mathematical Sciences, Queensland University of Technology; Infection Management Services, Princess Alexandra Hospital; and The Prince Charles Hospital. ARC Linkage projects support collaborative research development projects between higher education organizations and other organizations, including industry, to enable the application of advanced knowledge to problems (Australian Government, 2007).

This research was undertaken as part of a sub-project with the objective of developing new statistical tools for surveillance and monitoring of adverse events in hospitals which adjust for differing patient and operational risks. The motivation was that there is a pressing need to better utilize the human and capital resources and minimise adverse outcomes for patients. The project partners take the view, expressed by the Australian Council for Safety and Quality in Health Care (AC-SQHC, 2002), that the emphasis should be on improving the systems rather than blaming individuals. The philosophy that change rather than blame will foster quality improvement within health care institutions was clearly articulated by the project partners in letters (Morton, 2002a,b) to medical journals.
When proposing the guidelines for this research degree, the project partners noted that risk adjustment was an important aspect of monitoring adverse events. Patient risk adjustment techniques needed to be used in order to make some allowance for what might be difficult surgical cases or the general level of health of the patients. Risk adjustment methods could have various goals including a measure for costing hospital treatment, assessing health status for insurance payments, or indicating severity of illness for clinical purposes, such as the Acute Physiology And Chronic Care (APACHE) system for intensive care patients which involves over 100 measurements for a patient. Risk adjustment is also undertaken for schemes monitoring the number of adverse events for patients undergoing surgery. At the time the guidelines were set, the practice was to seek complex risk adjustment to enable comparisons between institutions, as Austin et al. (2001) did for example. However, the project partners considered that risk adjustment methods based on a few key variables were likely to provide adequate predictions of the probabilities of adverse outcomes when a single hospital was examining its own data. They noted that such a method had recently been implemented by the British Heart Foundation for myocardial infarction mortality (Lawrance et al., 2001). Thus, one of their research tasks was the identification of new risk adjustment techniques that would be suitable for schemes monitoring the number of adverse events for patients undergoing surgery.

The project partners noted that there were several challenges in terms of statistical data analysis and modelling for adverse event data. Some of the issues identified were that potential denominators for adverse event data are not necessarily the appropriate exposure variables to construct the comparisons, denominator variables may be subject to error, conventional modelling of clustered adverse events using the Poisson distribution is inappropriate because the data tend to be overdispersed, and the frequency of certain adverse events, such as death, may vary from approximately 15% for patients admitted to intensive care units to 1–2% for patients undergoing cardiac surgery. In the case of the rarer events, it may be more appropriate to consider the number of non-adverse events between adverse events, giving a discrete waiting time distribution.

In their guidelines for the candidate, the project partners noted that one
approach is to use sequential methods to monitor adverse events. They noted that sequential monitoring tools included control charts, such as Cumulative Sum (CUSUM) charts, $P$-Charts, and Exponentially Weighted Moving Average (EWMA) charts described in textbooks for statistical quality improvement (Montgomery, 2001; Ryan, 2000), scan statistics, and change-point models, but all need modification for appropriate denominator and risk adjustment. Another general approach that the guidelines flagged for investigation was to use Bayesian hierarchical models, for example, Weir and Pettitt (2000), so as to incorporate the questions of denominators, clustering, and variation in the probability model. Hay and Pettitt (2001) used hierarchical Bayesian techniques to explore the relationship between the incidence of *Klebsiella pneumoniae* and the use of the antibiotic class, 3rd generation cephalosporins, in the Princess Alexandra Hospital. Their rationale was that, because risk adjustment schemes may have as their focus either within-hospital standardization, for the explicit purpose of in-house process improvement, or between institutions, so that comparisons between them may be made, hierarchical Bayesian models might be developed to deal with the variety of data structures and to take account of the inherent variability so that analyses are useful to each hospital and to health care administrators.

According to Mohammed (2006), when statistical process control (SPC) was introduced into health care, its objective was confined to the role of judgmental monitoring, usually to discover poor performance. However, Thomson and Lally (2000) state that the use of health care data for judgment about performance is a tricky business. Mohammed (2006) compares and contrasts the philosophy for the application of statistical process control to health care, where researchers wish to place SPC in a statistical modelling paradigm in which failure to address statistical issues, such as multiple testing, is regarded as a serious omission with its use in the industrial context. He notes that SPC is used in industry as a way to understand and improve the underlying process and supports that statement with the following quote from Tukey (1946):

“A state of control, in Shewhart’s basic sense, is reached when it is uneconomic to look for assignable causes, and experience has shown that the compactness of distribution associated with a good control chart implies this sort of control.
It does not imply randomness! There seems to be two reasons for this common confusion between control and randomness. First, bringing a process in-control almost inevitably brings it measurably nearer randomness . . . Second, randomness is easily discussed in a simple model by a mathematical statistician, while economic control requires a complex model without (to date) any interesting or beautiful mathematics.”

Clearly, the philosophy of the ARC Linkage project partners is that SPC should be used in health care as a way to understand and improve underlying processes. They also identify risk adjustment of adverse outcomes as an important problem for SPC in health care. It is unimportant in most industrial production contexts because variability in the input of raw materials to industrial processes may be eliminated as it is usually special cause variation (Montgomery, 2001), but variability in the input of patients to medical process is clearly common cause variation. The research for this thesis focuses on models for risk adjustment and the effect of variable risk adjusted probabilities of expected outcomes on the performance of control charts. The particular methodology for the risk models used in this thesis is, in the main, logistic regression (Agresti, 2002, Chapter 5) and control chart studies are presented on risk adjusted CUSUM schemes (Steiner et al., 2000).

The literature review in Chapter 2 provides an overview of control charts used in health care and an in-depth review of risk scores to adjust the expected probability of mortality of trauma victims. A further in-depth review of the risk scores to predict the expected probability of mortality of cardiac surgical patients in Chapter 3 shows that risk factors for mortality following cardiac surgery may be classified into eight so-called meta-factor categories and that an increase in the number of covariates for a risk score does not result in a corresponding increase in the ability of the model to discriminate between the risks of patients’ mortality. Issues associated with the use of report cards for the comparison of institutions’ performances are discussed in Chapter 4. The debate about the use of report cards to monitor mortality rates following cardiac surgery in New York State is reviewed and trend analysis is suggested as an alternative method for reporting the mortality rates. There are also analyses of some data for the predicted probabilities and observed mortality outcomes of patients following their admission to one of
eleven hospitals in north-east England. The results show that the identification of “outlier” hospitals depends, to some degree, on the analytical method used. CUSUM schemes and, in particular, risk adjusted CUSUM schemes are discussed and analysed in Chapter 5. Methods used for assessing the performance of control charts are discussed and in-control and out-of-control average run lengths (ARLs) are chosen as the measures to assess the performance of risk adjusted CUSUM schemes, where the population of patients’ risk profiles varies. The chapter concludes with a proposal to stabilize the ARLs by constraining the step from the observed probabilities of mortality for in-control processes to the observed probabilities of mortality for out-of-control processes, to which the risk adjusted CUSUM scheme is tuned, to be no less than some minimum value. The effects of uncertainty at the patient level and in estimating risk model parameters on the performance of risk adjusted CUSUM schemes are explored in Chapter 6. The ARLs of the risk adjusted CUSUM schemes assessed in Chapters 5 and 6 were approximated using simulation because the approximations found using the Markov chain property of CUSUMs proved to be unstable. An alternative method of calculating the Markov chain transition probabilities to stabilize the resulting approximations of the ARL is proposed for patient populations risk adjusted by discrete finite risk models in Chapter 7. Summaries of the investigations undertaken are given in Chapter 8. The thesis concludes with some suggestions for further research for implementation of SPC in health care.
This chapter reviews risk adjustment and control charts for quality improvement in the health care setting under three broad topics. They are the use of standard control charts for monitoring medical processes, the development and use of control charts to monitor adverse events in health care, and the development of risk models for mortality in trauma victims. Some of the literature on these topics is reviewed in Sections 2.1, 2.2, and 2.3, respectively.

2.1 Standard Control Charts

The standard control charts considered in this section are those that are described in introductory textbooks on statistical process control for industrial processes. This section provides examples of their use for monitoring quality characteristics of clinical and administrative medical processes given in Morton (2005). Section 2.1.1 is concerned with control charts for monitoring quality characteristics that are expressed in terms of numerical measurements. In Section 2.1.2, the quality characteristics are measured as counts and, in Section 2.1.3, the measures are proportions.
2.1.1 Control Charts for Numerical Data

Shewhart Control Charts
Quality characteristics expressed as numerical measurements are often assumed to have a normal distribution with mean $\mu$ and standard deviation $\sigma$. Suppose the mean of a sample of size $n$ of the measurements is $\bar{X}$. If one assumes normality, then the Central Limit Theorem is stronger than result to obtain the sampling distribution. That is, $\bar{X}$ is also assumed to be normally distributed with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$. A control chart using samples from a process to monitor its average is called an $\bar{X}$ chart and the process variability may be monitored with either an $S$ chart, to control the standard deviations of the samples, or an $R$ chart, to control the range of the samples. A description of these charts and instructions for their construction may be found in Montgomery (2001, Chapter 5). Morton (2005) recommends the use of $S$ charts to control process variability. Such charts can be useful for displaying aspects of resource use, such as ward occupancy and pharmacy utilization.

Shewhart $I$ control charts monitor the means of individual numerical measurements, which are assumed to be normally distributed. Montgomery (2001) gives details of their construction, which are similar to those for $\bar{X}$ charts. Morton (2005) notes that it is important that the data are approximately normally distributed and recommends testing for normality with the Shapiro-Francia $W'$ test (Altman, 1991).

The $I$ charts may be used for monitoring the number of tests required to achieve a positive result in blood culture specimens and length of stay (LOS) in hospital for more than one month (Morton, 2005). LOS distributions are highly skewed so control limits were set using “relative to an identified distribution” (RIDIT) analysis (Bross, 1958). Another possible application for $I$ charts is the surveillance of antibiotic usage. This may be accomplished by monitoring the number of defined daily doses (DDDs) of an antibiotic prescribed in a period of interest, for example, one month (Morton, 2006). However, he found that the data are typically highly variable and that a better option was the EWMA chart.
EWMA Charts

Montgomery (2001, Chapter 8) describes the construction of EWMA control charts. EWMA charts are used for monitoring the process mean and are very robust to the normality assumption. If the weights and control limits are appropriately chosen, they will perform very well against normal and non-normal distributions. Borror et al. (1999) showed that the EWMA chart can be designed so that it is robust to the normality assumption; that is, so that the average run length to a signal when the process is in-control (ARL₀) is reasonably close to the normal-theory value for both skewed and heavy-tailed symmetric non-normal distributions. In contrast, moderate non-normal distributions had the effect of greatly reducing the ARL₀ of Shewhart $I$ charts, increasing the rate of false alarms.

2.1.2 Control Charts for Count Data

Shewhart $C$ and $U$ Control Charts

In the industrial context, Shewhart $C$ charts are used to monitor number of non-conforming items in samples of constant size $n$. The number of non-conforming items from an in-control process is assumed to follow a Poisson distribution with parameter $c$, if the expected number of non-conforming items is known, or with an estimated parameter $\bar{c}$, if the number non-conforming is unknown. The estimate, $\bar{c}$, may be the observed average of non-conformities in a preliminary dataset. If the size of the samples is not constant, the number of non-conforming item may be monitored using a Shewhart $U$ chart, where the average number of non-conformities per inspection unit are monitored. Details of the use and construction of $C$ and $U$ charts are given in Montgomery (2001, Chapter 6).

The Shewhart $C$ charts have been used in the clinical context to monitor monthly counts of bacteraemias where there is no reason to doubt the bacteraemias are independent and it is assumed the counts follow an approximate Poisson distribution (Morton, 2005). Monitoring monthly counts using $C$ charts is appropriate if the monthly number of occupied bed days in the hospital is relatively stable. Where there is considerable variation in a hospital’s occupancy, the appropriate monitoring tool is the $U$ control chart which monitors the number of bacteraemias
or infections per 1,000 occupied bed days.

Although standard $C$ or $U$ charts may be used to monitor independent events such as surgical site infections or bacteraemias, Morton et al. (2001) note that such charts may not be adequate for surveillance of colonizations or infections with multiply resistant organisms (MROs), such as extended spectrum beta lactamase (ESBL) *Klebsiella pneumoniae* or methicillin resistant *Staphylococcus aureus* (MRSA). In such circumstances, the variance of the counts may exceed their mean and false positive signals from the standard control charts may be excessive. The false positives may be reduced by setting the control limits of the Shewhart charts using the negative binomial distribution. For example, Morton (2005) found that the counts of “monthly MRSA new isolates” in the (Princess Alexandra) hospital between July 1995 and December 1996 had a variance far greater than the mean and he recommended the use of the negative binomial distribution for setting appropriate control limits.

It is of interest that Cooper and Lipsitch (2004) used monthly counts of non-duplicate clinical isolates of MRSA from 41 intensive care units (ICUs) at U.S. hospitals to show that a simple Poisson model was inadequate. They concluded that a structured hidden Markov model based on a continuous time epidemic model gave the best fit to the ICU data when compared to the Poisson models, but that the hidden Markov model may not be appropriate for larger hospital populations made up of several interacting units. Drovandi (2006) confirmed that the structured hidden Markov model provided the best fit in terms of the Akaike Information Criterion (AIC).

*EWMA Control Charts*

There have been extensions and variations of the basic EWMA control chart. Montgomery (2001) gives some examples as the fast initial response feature, monitoring variability using an exponentially weighted mean square error (EWMS) control chart, using the EWMA as a predictor of process level, and the EWMA for Poisson data. The EWMA chart has been used in the medical context to display small persistent changes in the mean value of bacteraemias and MRSA colonizations (Morton, 2005) and for syndromic surveillance (Burkom, 2006).
2.1.3 Control Charts for Proportion Data

Shewhart Control Charts

In industry, the Shewhart $P$ control chart is used to monitor the fraction of items that do not conform to one or a number of standards. Its design and an example showing how it may implemented to monitor the output of an industrial production process is given by Montgomery (2001, Chapter 6). Morton (2005) gives examples of its use to monitor complications that occur in the medical context. It was used to monitor the proportion of patients whose LOS exceeds a certain percentile for the relevant diagnosis related group (DRG), the proportion of surgical patients with surgical site infections, mortality in large general purpose intensive care units in public hospitals, which is about 15%, and the proportions of various hospital records that are completed correctly. Control limits may be set using the normal approximation of the binomial distribution but, if the number of trials is small or $\bar{p} \sim 0.1$, the Camp-Poulson approximation (Gebhardt, 1968), derived from the F-distribution, is a more appropriate approximation.

If the expected proportion is less than 10%, Morton et al. (2001) proposed that the Poisson approximation to the binomial distribution be used for data analysis. The data are grouped into blocks such that there is one expected complication per block, if the process is in-control. Then any special cause variation may be detected by monitoring the complications with Shewhart $C$ control charts. This monitoring scheme has been superseded by count data CUSUM charts.

CUSUM Charts to Monitor Small Proportion Data

Morton (2005) found that control charts such as the EWMA or CUSUM charts are better than Shewhart charts for detecting small step changes in any quality measure being monitored. If that measure is a proportion of less than 10% for the in-control process, count data CUSUM schemes may be used to detect step shifts in the proportion. Beiles and Morton (2004, Figures 6 and 7) illustrate the CUSUM methods with an example where the rate of surgical site infections are monitored to detect a shift from a surgical site infection rate of approximately 3% for the process in-control to a rate of approximately 6% for the process out-
of-control. In the informal CUSUM graph the cumulative number of infections is plotted against the number of operations undertaken. An out-of-control signal occurs if the CUSUM graph crosses above the line given by

\[ \text{Infections} = 0.06 \times \text{Operations}. \]

In the CUSUM test, the surgical site infection data are grouped into blocks of 30 operations so that the expected number of infections is approximately one per block. He then implements a CUSUM scheme which is equivalent to the decision interval CUSUM scheme for Poisson data given by Hawkins and Olwell (1998, Chapter 5.4). Ismail et al. (2003) developed a monitoring scheme based on the scan statistic which they found sensitive in detecting changes in surgical site infection rates that might not be detected using standard control chart methods such as the CUSUM test.

### 2.2 Control Charts for Monitoring Adverse Events in Health Care

The literature review in Section 2.2.1 concerns control charts that are designed to monitor adverse events that occur infrequently. Risk adjusted control charts are reviewed in Section 2.2.2, and charts to monitor multiple institutions or clinicians are considered in Section 2.2.3. This section concludes with a discussion of some of the issues for evaluating the performance of controls charts in Section 2.2.4.

#### 2.2.1 Charts to Monitor Rare Events

*Charts based on the Geometric and Negative Binomial Distributions*

Ryan (2000, Section 6.1.6) states the use of a \( P \) chart is inadvisable if \( p \) is very small. In such circumstances, he recommends control charts that plot the number of events until \( k \) non-conforming items are observed. If \( k = 1 \), the appropriate distribution is the geometric distribution and, if \( k > 1 \), it is the negative binomial distribution.
There are issues with calculating control limits of such charts because the negative-binomial and geometric distributions are extremely skewed. Ryan (2000) recommends unbiased control limits be set based on analysis of the control charts’ run length distributions proposed by Acosta-Mejia (1999). Yang et al. (2002) considered the effect of uncertainty on estimation of the proportion of non-conforming items from a large-volume, high-quality, in-control production process. They concluded that its effect on the average run length of a geometric chart is only mild unless the sample size is (relatively) small or there is a large process improvement, where a small sample size is less than 10,000 items and the proportion of non-conforming items ranges from $10^{-4}$ to $5 \times 10^{-3}$.

A $g$ EWMA control chart, based on inverse sampling from the geometric or negative binomial distribution, was developed by Benneyan (2001) to monitor the number of events between adverse outcomes, where the probability of an adverse event is low. It is simple to use and can exhibit significantly greater detection power than conventional binomial approaches, particularly for infrequent events and “low” detection rates. Benneyan (2006, Figure 7) uses a $j$-binomial standardized version of the $g$-EWMA control chart to monitor surgical site infections where the patient population is stratified into $j$ categories where the risk adjustment is according to the four National Nosocomial Infection Surveillance (NNIS) risk categories. The NNIS index may be calculated by scoring one point for the following three risk factors: (i) contaminated or dirty-infected surgery, (ii) American Society of Anesthesiologists’ score higher than 2, and (iii) duration of surgery longer than $T$ hours, where $T$ depends on the surgical procedure (Delgado-Rodriguez et al., 2006).

The Sets Method
The sets method was proposed by Chen (1978) as an alternative to CUSUM schemes used for surveillance of the number of newborns with a specific congenital malformation. A set is the number of normal births between each infant born with a malformation and signals an alarm if a sequence of sets appears such that each set is below a certain size. Woodall (2006) notes that many authors have studied the sets methods and made various modifications. Chen (1978) proposed
the sets method as a surveillance system with simpler calculations than those for the CUSUM scheme. However, Sego (2006) used the steady-state average run length (Hawkins and Olwell, 1998, Section 3.3.4) to show that the performance of the Bernoulli CUSUM scheme is superior. The Bernoulli CUSUM scheme is as described by Sego (2006, Section 3.4). The CUSUM weight (Steiner et al., 2000) is calculated from the Bernoulli likelihood ratio in Wald’s sequential probability ratio test.

**Risk Adjusted Survival Time CUSUM**

The risk adjusted survival time (RAST) CUSUM was proposed by Sego (2006) for monitoring clinical outcomes where the primary endpoint is a continuous, time-to-event variable that is right censored. Risk adjustment is achieved using accelerated failure time (AFT) regression models based, in part, on their discussion by Klein and Moeschberger (1997). If the survival times are assumed to follow a log-logistic distribution, the performance of the RAST CUSUM scheme was superior to that of the risk adjusted CUSUM scheme (Steiner et al., 2000), especially when the fraction of censored observations was not too high. If the censoring fraction for the in-control process was 97.5%, the performances of the two CUSUM schemes were comparable.

### 2.2.2 Risk Adjusted Charts

**Variable Life Adjusted Display**

The variable life adjusted display (VLAD) proposed by Lovegrove et al. (1997) plots the cumulative difference between the expected and observed cumulative deaths, \( \sum_t p_t - \sum_t y_t \), where \( p_t \) is the expected probability of mortality given by an appropriate risk model and the observed outcome \( Y_t \) of patient \( t \) is a binary indicator. Grigg and Farewell (2004a) note that the VLAD chart represents an intuitively useful way to represent performance over time, but the plot is not the most natural from which to determine if and when an alarm should be raised.

**Cumulative Risk Adjusted Mortality chart**
The cumulative risk adjusted mortality (CRAM) chart also plots the cumulative difference $\sum p_t - \sum y_t$. Poloniecki et al. (1998) assumed the number of deaths for an in-control process followed a Poisson distribution with mean called the adjusted estimate calculated from a so-called in-control performance ratio, which they define, and the expected probabilities of patient mortality. Control limits were set according to $\chi^2_1$ statistics used in tests for any change in the performance ratio at the 0.01 significance level after each death. In assessing the chart’s performance, Steiner et al. (2000) note that consideration of a false alarm rate does not make sense. Their reasons for that statement are that there is no formal adjustment for multiple testing and the control limits effectively change over time because there is continual updating of the in-control performance ratio. Sismanidis et al. (2003) used simulation to approximate in-control and out-of-control run lengths, which were defined as the number of failures (deaths) after the first trial.

**Risk Adjusted Shewhart $P$ Charts**

Cook et al. (2003) proposed a risk adjusted Shewhart $P$ chart to monitor standardized mortality rates in an intensive care unit. The authors state the control limits were dependent on admission numbers, the distribution of the expected probabilities of patient mortality $p_t$ predicted by the risk score, and the assumption that the number of deaths in a sample of $n$ patients is modelled by a normal distribution with mean $\sum p_t$ and variance $\sum p_t(1 - p_t)$. Grigg and Farewell (2004b) state the normal approximation is thought to be reasonable provided the distribution of the patient probabilities is tight and non-skew, but if it is flat or highly skewed the ARL should perhaps be checked by simulation.

Hart et al. (2003) proposed six risk adjusted Shewhart $P$ charts to monitor health care performance. Two of their charts effectively monitor standardized mortality rates and another two are similar to CRAM or VLAD charts. They also proposed a system where two $P$ charts would be used in tandem to monitor the observed outcomes and the expected probabilities predicted by the risk score. Monitoring the observed outcomes was suggested to answer the question as to whether the observed rates are stable over time and monitoring the expected outcome rate was suggested to answer the question as to whether the expected rates
are unstable over time. Hart et al. (2003) state that both these are of interest because it is important to know both the observed and expected mix of patients for planning of resources and to check the integrity of the data.

*Resetting Sequential Probability Ratio Test*

Woodall (2006) states the resetting sequential probability ratio test (RSPRT) chart was proposed by Morton and Lindsten (1976) to monitor the rate of Down’s syndrome. Grigg et al. (2003) described the RSPRT scheme as a sequence of SPRTs (Wald, 1947) with an absorbing boundary at $b$ and a reflecting boundary at $a$, where $a$ and $b$ are real numbers. They note that the CUSUM scheme is a special case of the RSPRT, where $a = 0$ and $b = h$ are the resetting barrier and the decision threshold, respectively. In comparisons of out-of-control ARLs of the CUSUM scheme and RSPRT charts with varying $a$ and $b$, Grigg et al. (2003) found that, if the process went out-of-control after 1,900 observations, the CUSUM scheme had the shortest average run length to a signal when the process is truly out-of-control (ARL$_1$). They concluded that the RSPRT chart is more likely to build up “credit”, or “inertia” as it is termed in industrial SPC (Woodall, 2006), than the CUSUM scheme. A risk adjusted RSPRT chart to monitor mortality in medical processes, such as adult cardiac surgery, was proposed by Spiegelhalter et al. (2003a).

*Risk Adjusted CUSUM Schemes*

The risk adjusted CUSUM scheme proposed by Steiner et al. (2000) is described and assessed in Chapter 5. Their purpose was to monitor adverse outcomes of medical procedures for which the patients’ expected probabilities of an adverse outcome is variable. Woodall (2006) noted that the first risk adjusted Bernoulli CUSUM scheme, where the expected probability of the outcome was modelled using logistic regression, was proposed by Lie et al. (1993) for the surveillance of Down’s syndrome among newborn. It was used to monitor if environmental agents, such as X-rays, increased the risk of the syndrome above that attributable to maternal age, rather than as part of a process improvement program.
2.2.3 Comparing Institutional Performance

Report Cards
Report cards, which frequently employ league tables, are reviewed and assessed in Chapter 4. Typically, a report card lists the risk adjusted mortality rates (RAMR) for treatments and procedures undertaken by all institutions managed by a health-care administration. Any RAMR that is “significantly” different—in some statistical sense—from the average is usually flagged or, in the case of a league table, the institutions are ranked in order of their RAMRs.

Comparison Charts
Adab et al. (2002) adapted a graphical technique proposed by Mosteller and Tukey (1949) to plot the square root of the number of heart attack victims admitted to each of 37 large acute hospitals in England during 1988-9, \( \sqrt{y_i} \), \( i = 1, \ldots, 37 \), against the square root of the number of patients admitted to each hospital, \( \sqrt{n_i} \). The square root transformation allows constant control limits because \( \sqrt{y_i} \) has an approximate mean of \( \sqrt{n_i \bar{p}} \), where \( \bar{p} \) is the expected average mortality rate for in-control processes in any coronary care unit, and a constant variance of approximately \( \frac{n}{4} \) (Mosteller and Tukey, 1949). The rationale for using comparison charts is that there is an explicit assumption that all hospitals are part of a single system and have similar performances, in contrast to a league table which implicitly assumes that there is a performance difference between hospitals (Adab et al., 2002). Presentation of data using comparison charts rather than league tables tends to reduce over-investigation of unusual performance because health service decision-makers identify fewer outliers and have a reduced tendency to request further information (Marshall et al., 2004b).

Funnel Plots
Spiegelhalter (2002) proposed the use of funnel plots for the graphical display of performance indicators of institutional performance as an alternative to comparison charts because the square root transformation appears unintuitive, obscures the event rates, and leads to rather approximate controls limits. Spiegelhalter
(2005) expresses a preference for the random effects model for estimating the degree of overdispersion and adjusting the control limits, as this model seems best to mimic the belief that there are unmeasured factors that lead to systemic differences between the true underlying rates in institutions. Spiegelhalter also notes that the funnel plot is a standard tool (used) within meta-analysis as a graphical check of any relationship between effect estimates and their precision that may suggest publication bias. He then provides examples of their use for comparing clinical outcomes and provides full details for their construction. They have many advantages; for example, (a) the axes, where the proportion of adverse outcomes are plotted against the number of cases, are readily interpretable, (b) there is no spurious ranking of institutions, and (c) there is a clear allowance for additional variability in institutions with small volume. Two limitations are that the only allowance for multiple comparisons is the use of small $p$-values for control limits and there is no formal allowance for “regression to the mean” in which extreme outcomes are likely to be a run of bad luck.

*Use of Bayesian Hierarchical Models*

Berlowitz et al. (2002) provide an example where formal allowance for regression to the mean changes the number of outliers when they profiled nursing homes. If they used the “standard” statistical technique to calculate the risk adjusted rate of pressure ulcer development, they identified twelve of 108 nursing homes with a higher than expected rate and three with a lower than expected rate of pressure ulcer development. However, if the analyses was undertaken using a Bayesian hierarchical model, only two of the 108 hospitals were identified with a higher than expected rate and none with a lower than expected rate of pressure ulcer development.

*Control for False Discoveries in Multiple Monitoring*

Following the crimes of Dr Harold Shipman (Smith, 2002), it became clear that there was little monitoring of the clinical governance of general practitioners in the United Kingdom. Aylin et al. (2003) investigated the use of multiple CUSUM charts to monitor the number deaths of patients belonging to each general practice
or individual general practitioner. They identified overdispersion in the data for
the process in-control as a cause of unnecessarily high false-alarm rates. They
also found that the CUSUM design process—of appropriate choices of an alarm
threshold and in-control and out-of-control death rates—for the tradeoff between a
tolerable number of false alarms and the timely detection of out-of-control processes
was an issue if CUSUM schemes were used for multiple monitoring of the clinical
governance of medical practitioners.

Marshall et al. (2004a) describe a model to adjust for overdispersion which uses
the false discovery rate (Benjamini and Hochberg, 1995) to control the number of
false alarms by specifying the expected number of processes out of control. Woodall
(2006) concluded that the assumptions were restrictive. Grigg and Spiegelhalter
(2006) agreed that the assumptions made by Marshall et al. (2004a) to control for
false discoveries are restrictive. Their method requires prior specification about
the number of expected units out-of-control over a fixed period of analysis. Grigg
and Spiegelhalter (2006) suggested an alternative analysis where the expected pro-
portion of units out of control at any instant of time is prespecified, but the state
of any one unit is transitory.

Rice and Spiegelhalter (2006) proposed a new graphical tool which identifies
outliers where there is the multiple comparison problem. They noted the suggested
procedure quantifies outlier status based on the estimated false discovery rate.

2.2.4 Performance Evaluation Issues

Woodall (2006) identifies several performance evaluation issues. First, he notes
that, in industrial quality control, it has been beneficial to carefully distinguish
between the Phase I analysis of historical data and the Phase II “on-line” mon-
itoring stage, but there is often no clear distinction between the two phases in
health-related control charting. Steiner (2006) agrees that there is a lack of ap-
preciation in medicine for the distinction between Phase I and II of the control
chart and gives an example (Jones et al., 2001) that illustrates the effect of estima-
tion error on the use of control charts. However, Grigg and Spiegelhalter (2006)
note that it is likely that the notions of Phase I and II in health-related control
charting are often blurred because Phase I conditions are those of a controlled experiment which, except in the restricted context of randomized clinical trials, is very difficult and expensive to manage because, they suggest, medical processes are dynamic. Burkom (2006) also considers that medical processes are dynamic and he envisages a steady state where Phase II monitoring is applied but can be automatically adjusted as the result of continued checking of distributional and correlational assumptions and of sample data quality.

A second issue is the use of $\text{ARL}_1$ only to evaluate the out-of-control performance of control charts. Performance measures to assess control charts used for monitoring medical processes are reviewed in Section 5.2. The decision to use $\text{ARL}_1$ for the evaluation of risk adjusted CUSUMs in the research undertaken for this thesis was made because Hawkins and Olwell (1998, Section 2.2) note that the steady state $\text{ARL}_1$ of CUSUM schemes are necessarily smaller than the $\text{ARL}_1$ because the change could occur when the CUSUM is at some value greater than zero. They state the relative rankings of different CUSUMs are generally not greatly affected by the assumption about the starting point.

Another issue Woodall (2006) raises is overdispersion resulting when the variance of the response exceeds what would be expected under a specified model, for example, the Poisson model. He notes that many authors recommend that the issue of overdispersion in count data is overcome in some cases by replacing the Poisson model with the negative binomial model. However, there are many unresolved issues on how to best adjust control charts for overdispersion.

### 2.3 Severity Measurement Systems for Trauma Victims

This section is concerned with reviewing examples from the literature concerning “severity-of-illness” measures which are, if the definition by Iezzoni (1994c) is generalized, systems that quantify risks of the short-term outcomes for patients who have been diagnosed with medical conditions, undertaken medical treatment, or have undergone medical procedures. In the context of measuring outcomes to
monitor the quality of particular treatments or procedures, the reason for severity measurement systems is to adjust for differences in patients’ risks so that institutions or medical practitioners are not penalized for accepting riskier patients (Iezzoni, 1997). The stated rationale for risk adjustment is to remove one source of variation, leaving residual differences to reflect quality. The underlying assumption is that outcomes result from a complex mix of factors: such that patient outcomes equal effectiveness of treatments plus patient risk factors that reflect response to treatment plus quality of care plus random chance.

“Risk adjustment” controls for patient risk allowing isolation of quality differences, but the term is meaningless if the risk is not identified. Iezzoni (1997) observes that identical risk factors may have different relationships with different outcomes or may predict one outcome, but not another. The necessity for constructing a severity of illness measure that is specific to the particular outcomes of interest has resulted in a proliferation of risk scores in the literature. An internet site providing details of over 10,000 computations, formulae, surveys, or look-up tables useful in health care (Institute for Algorithmic Medicine, 2006) has listed risk score algorithms for the various outcomes for many types of patients. Some examples of outcomes, for which risk scores have been constructed, are:

- Mortality of trauma patients;
- Mortality of patients admitted to adult, pediatric, and neo-natal intensive care units;
- Mortality of hospital patients (see Iezzoni (1994b) for examples);
- Co-morbidities in patients, for example the Charlson comorbidity index (Charlson et al., 1987);
- Mortality and morbidity of patients following stroke;
- Mortality and morbidity of asthma patients;
- Acquisition of coronary artery disease;
- Cardiac arrest following general surgery;
• Mortality of patients following admission to a coronary care unit with acute myocardial infarction;

• Mortality of patients following cardiac surgery;

• Complications following anaesthesia and surgery;

• Adverse outcomes for patients with blood disorders;

• Deep vein thrombosis and anti-coagulant related bleeding;

• Complications associated with acute and chronic pancreatitis

• Type 1 and Type 2 diabetes mellitus and hypoglycaemia in patients who use insulin;

• Renal failure;

• Infiltrating ductal carcinoma for breast cancer patients;

• Spontaneous preterm deliveries for pregnant women;

• Complications with a foetus or neonate;

• Complications during labour and delivery;

• Severe depression;

• Pressure sore development;

• Bacteraemia and sepsis (including Staphylococcus aureus bacteraemia);

• Anthrax rather than some other infection;

• Surgical site infections; and

• Dental caries \(^1\).

\(^1\)Anderson et al. (2002) defines dental caries as a tooth disease caused by the complex interaction of food, especially starch and sugars, with the bacteria that form dental plaque. The term also refers to the tooth cavities that result from the disease.
Iezzoni (1997) found that hospital severity measures failed to fully explain the variation in unadjusted death rates across hospitals. She and other researchers were unable to establish the truth or otherwise of the assumption that risk adjustment isolates a residual quantity; namely, quality of care differences across hospitals. She concluded that a definitive answer was unlikely any time soon.

The use of severity measures as decision making tools remains controversial. At the time this chapter was being prepared, Aldhous (2007) reported the British government was rolling out a treatment and research program called the “Dangerous People with Severe Personality Disorder” (DSPD) program. Offenders would be identified as DSPD, and possibly incarcerated indefinitely, if they had a PCL-R score (Hare, 1991) of 30 or above, a PCL-R score of 25–29 and a diagnosis of at least one personality disorder other than antisocial personality disorder, or two or more personality disorders, but there was widespread criticism of the definition of DSPD. Robert Hare, who constructed the PCL-R score, said that it was an amalgam of clinical disorders that made no sense as a diagnostic category and Jeremy Coid, a forensic psychiatrist, said that, although it is possible to assess someone’s likelihood of committing further violence, there is no accepted way of establishing that an offender’s dangerousness is linked to their personality disorder.

Despite the vagueness of risk adjustment, Iezzoni (1997) provides compelling reasons for continuing its use in quality programs for medicine. The major reason is that, however imperfect, there is no other way to begin a productive dialogue with physicians and other clinicians about using outcomes information to motivate quality improvement. Thus, in the context of the objectives of this thesis—surveillance and monitoring of adverse medical outcomes—it appears that risk adjustment should be incorporated in any tool to monitor adverse events. If there is no risk adjustment, a monitoring tool will not be accepted nor used by the healthcare workers for whom it was designed. Accordingly, this part of the literature review attempts to provide an understanding of the strengths and weaknesses of risk adjustment by reviewing in detail the risk scores for mortality of trauma victims.
2.3.1 Risk Scores for Triage of the Trauma Patient

History and Development of Triage Systems

In military medicine the word triage, which is derived from the French verb *trier*—
to sort out, refers to the classification of casualties of war and other disasters according
to gravity of injuries, urgency of treatment, and place of treatment (Anderson et al., 2002).
In their historical perspective of mass casualty incidents, Nocera and Garner (1999) note
that, prior to 1792, wounded soldiers were left where they fell until the end of the battle.
After completion of the battle they were located and then evacuated according to rank,
with deceased nobles being attended to before wounded common soldiers. In 1792, Dominique Jean Larrey,
Surgeon General of Napoleon’s Army of the Rhine, introduced his innovation of
rescuing casualties during battle with a dedicated corps and purpose built wagons,
the *ambulance volantes*. After evacuation to the rear of the battlefield, the order of
initial care of the wounded was based on whether there was threatened loss of life
or limb. Those with minor injuries were given first aid and sent back into battle
(Hoyt et al., 2004).

According to Nocera and Garner (1999), the next innovation, described in 1846
by Dr John Wilson, British naval surgeon, was to prioritize the treatment of life
threatening haemorrhage and to classify combat injuries into slight, serious, and
fatal. In a more modern two-tiered triage scheme, the United States medical corps,
who provide first echelon care to battlefield casualties, first classify the patient ac-
cording to the scheme proposed by Dr Wilson then sort the serious casualties,
who require transport to a field or base hospital, into triage categories of urgent,
immediate, and delayed, where the injury categories are life threatening requiring
immediate treatment and resuscitation, life threatening requiring treatment within
two hours, and not immediately life threatening nor compromised by delayed treat-
ment, respectively (Llewellyn, 1992).

Triage systems lead to improvement in casualty survival. For example, the
combination of triage, advanced resuscitation, and rapid helicopter evacuation of
casualties in the Vietnam War contributed to reduced casualty rates of 1% com-
pared with 4.7% observed during the Second World War (Nocera and Garner,
Objectives of Triage

In trauma medicine the objective is to match the patients with the optimal resources necessary to adequately and efficiently manage their injuries and the challenge is to correctly identify injury victims in need of a designated trauma centre. According to Hoyt et al. (2004), the objective and challenge will be met with a triage system that transports all seriously injured patients to the appropriate medical facility, alerts the trauma centre to facilitate treatment beginning on arriving at the hospital, defines the “major trauma victim” by analysis of patients’ injuries after they have been identified, and monitors the rates of overtriage and undertriage. Overtriage is defined as the false positive rate, that is, the proportion of patients transported to a designated trauma centre but not having suffered major injury and undertriage is the proportion of seriously injured patients triaged to a non-designated site (Terregino, 1998). There is some discussion of the consequences of over- and undertriage later in this section.

Trauma Scores for Triage of Single-Casualty Incidents

Typically, a trauma score for the triage of single casualty incidents is derived by summing the integer values assigned to its risk factors for mortality of trauma patients. Depending on how the integer values are assigned, a decision to transport the victim to a trauma centre is made if the score is above or below the decision threshold. The risk factors used to construct the trauma score are taken from one or a combination of anatomical, physiological, mechanism of injury, or comorbidity variables (Hoyt et al., 2004). Anatomical data are collected by physical examination which, in the field, is time consuming and has the potential to distract the examiner and the patient by focussing attention on fractures and soft tissue injuries and away from more subtle, life-threatening injuries. Consequently, sophisticated injury severity scores which use anatomical data are usually calculated after treatment is finished (Lefering, 2002) for use in quality programs. Injury severity scores are described in Section 2.3.2.
Physiological risk factors include the vital signs, such as blood pressure, heart rate, respiratory rate, and level of consciousness. Trauma scores that have been constructed in the past using only physiological factors include the Glasgow Coma Scale (Teasdale and Jennett, 1974), the Triage Index (Champion et al., 1980a), the Trauma Score (Champion et al., 1981), the Revised Trauma Score (Champion et al., 1989), the Mainz Emergency Evaluation Score (MEES) (Hennes et al., 1992), and the Cardiovascular Respiratory Score (CVRS) (Champion et al., 1981; Asensio et al., 1998).

An example of an anatomical risk factor is head injury. Some scores that use anatomical risk factors or both physiological and anatomical risk factors are CRAMS (Circulation, Respiration, Abdomen, Motor, and Speech) Scale (Gormican, 1982), Pre-hospital Index (Koehler et al., 1986), the Trauma Triage Rule (Baxt et al., 1990), Criteria to Identify the Unstable Patient After Trauma (Rhodes, 2002), Field Categories of Trauma Patients (American College of Surgeons Committee on Trauma, 1980), and Specific Injuries Identified During Field Categorization Warranting Transfer to a Trauma Centre (American College of Surgeons Committee on Trauma, 1980).

Other categories of risk factors used in trauma scores include mechanism of injury such as penetrating injuries, falls of more than four metres, and pedestrians struck by a car; comorbidities such as diabetes mellitus and intoxication with alcohol or other recreational drugs; environmental concerns such as exposure to heat or cold; and paramedic judgement (Hoyt et al., 2004). Examples of scores which include risk factors from these categories are the Trauma Index (Kirkpatrick and Youmans, 1971), the Revised Triage Scale and Checklist (Kane et al., 1985), Triage of the Automobile Trauma Patient Based on Vehicle Damage (Jones and Champion, 1989), Combining the Prehospital Index (PHI) and Mechanism of Injury (MOI) for Field Trauma Triage (Bond et al., 1997), New Jersey Standard Triage Criteria (Lavery et al., 2000), Venous Lactate Concentrations in the Triage of a Trauma Patient (Lavery et al., 2000), the Trauma Criteria of the Ohio State University Hospitals (Cook et al., 2001), Trauma Triage Rules with Mortality Risk Equation (Tinkoff and O’Connor, 2002), Patients for Whom Resuscitation is Futile (Salomone and Frame, 2004), and Triage Criteria for Possible Blunt Cardiac
Trauma Following a Fall from Height (Türk and Tsokos, 2004).

Trauma scores may be used to assess patients’ fitness for air evacuation. They usually provide a status score for the patient’s current clinical condition and a predictor score to estimate the patient’s potential for recovery (Institute for Algorithmic Medicine, 2006). Examples of triage scores for aircraft use are the SIMBOL Rating for Air Medical Evacuation (Williams and Schamadan, 1969), National Committee on Aeronautics (NACA) Score (Veldman et al., 2001), and Indications for Using Air Medical Transport of a Trauma Patient (Cole, 2002).

The proliferation of trauma scores for triage is probably motivated by a continuing quest for a “perfect” score (Terregino, 1998). In her review of trauma scores for field triage, Terregino (1998) noted that the goal of prehospital trauma indices is to get the right patient to the right hospital at the right time. She did not find any that were consistently better than the others in identifying major trauma victims or any that had been universally accepted.

Overtriage and Undertriage
In single-casualty incidents, the purpose is to classify the victims as major trauma patients needing trauma centre care or patients not requiring trauma centre care. The potential effect of undertriage on the patient is obvious and the adverse effect of overtriage is it is expensive and fatiguing for the hospital and staff (Senkowski and McKenney, 1999). Clearly, another possible effect of overtriage of patients who are not seriously injured is a potential block to the prompt treatment of major trauma victims.

Because of the uncertainty in identifying all major trauma victims, the American College of Surgeons states that a rate of 50% for overtriage of patients to a designated trauma centre may have to be accepted in order to maintain undertriage of patients to a non-designated centre at a rate of less than 10% (Terregino, 1998). In 1993, the College proposed guidelines for pre-hospital triage of trauma victims in which the decision to transport the patient to a trauma centre is made if there is any sign or symptom from the vital signs, level of consciousness, anatomy of injury, mechanism of injury, or patient characteristics indicating that the injury could be serious. Full details of the American College of Surgeons guidelines are
given by Terregino (1998, Table 6), Senkowski and McKenney (1999, Figure 4), or Hoyt et al. (2004, Table 4-4).

One method of controlling overtriage of patients to a trauma centre is secondary triage for a two tier personnel response (Phillips and Buchman, 1993). During the patient’s transport to hospital, a physician from the trauma centre is in contact with the pre-hospital team and decides if the patient requires ICU and/or operating room (OR) services. If the patient requires these services, the full response team treats the patient on arrival at the trauma centre, otherwise the modified response team provides treatment. The process is dynamic in that, if re-evaluation shows deterioration of the patient’s condition, a modified response may be upgraded to the full response. One analysis of two tiered response schemes has shown a sensitivity of 95% for predicting which patients will require ICU/OR services within 24 hours of admission to hospital while avoiding urgent trauma team mobilization in 57% of patients triaged for secondary response (Phillips and Buchman, 1993). In this instance, sensitivity is defined as the probability that a patient is identified for the full response given that the patient requires the full response. Another study showed two-thirds of the pediatric patients triaged for full response and 3% of those triaged for a modified response met the criteria for severe injury (Terregino, 1998). A third found savings on personnel, operating room, laboratory work, and protective wear of over $1,000 for each patient triaged for modified response (DeKeyser et al., 1994).

Trauma Scores for Triage of Multiple-Casualty Incidents

If there are multiple or mass casualties, it is probable established triage procedures will be modified so that limited available resources are used to provide medical care to the greatest number of patients (Hoyt et al., 2004). Triage becomes a process to rapidly identify casualties at the extremes of care. The purpose is to divert medical resources from those who will die or recover irrespective of the medical care they receive to critically ill casualties with a reasonable probability of survival (Nocera and Garner, 1999). In general, the triage systems for multiple-casualty incidents (MCI); for example, the Military Triage (Llewellyn, 1992), the Triage Sieve (Advanced Life Support Group, 1995), the Homebush Triage Standard Using Sim-
Triage Assessment and Rapid Transport (START) and Secondary Assessment of Victim Endpoint (SAVE) (Nocera and Garner, 1999), and CareFlight Triage (Garner et al., 2001); use physiological factors, such as level of consciousness and vital signs, to classify the victims as delayed treatment, immediate treatment, or unsalvageable. Triage systems for the MCI context are designed with a simple structure and based on normal daily operating procedures (Nocera and Garner, 1999). The rationale is that people are likely to perform better in the stress of a MCI if procedures for disaster response are kept as near as practical to their normal routines.

In the event of a major disaster such as a catastrophic earthquake, it is possible the goals of the triage of multiple casualties must be modified because there may be multiple and widespread disaster scenes, infrastructure may be damaged, and outside assistance may not be available to the local health care providers for two to three days (Benson et al., 1996). Because available medical resources are likely to be limited in such austere conditions, a major disaster plan will optimize use of resources by implementing a secondary triage system, called SAVE, after the initial triage, called START, to classify victims as delayed care, immediate care, or unsalvageable. The purpose of SAVE is to identify the patients with poor prognoses or those whose outcomes are unlikely to be affected by the treatment available at the scene of the disaster.

### 2.3.2 Trauma Scores for Quality Improvement

*Scores Based on Patterns of Anatomical Injury*

The original Abbreviated Injury Score (AIS) was an anatomical scoring system introduced in 1969 (Gunning and Rowan, 1999; Trauma Scores, 1969) with the purpose of providing safety data to automotive design engineers (Senkowski and McKenney, 1999). The severity of each of the injuries a patient sustained was rated on scale of 1 to 9 with non-fatal injuries rated 1 to 5. The guidelines for the original system, which defined 73 blunt injuries and assigned a consensus-derived severity score to each injury (Committee on Medical Aspects of Automotive Safety, 1971), were published in 1971. Greenspan et al. (1985) published revised guidelines (AIS-
CHAPTER 2: LITERATURE REVIEW

Injury Severity Scores
The original Injury Severity Score (ISS) (Baker et al., 1974) is an empirically derived score which provides a mechanism for summarizing multiple injury codes of the AIS into a single score and using that score to compute expected probabilities of mortality for patients classified by age into categories of less than 50 years, 50 to 69 years, or 70 years or more (Osler et al., 1996, 1997). The ISS is a risk-adjustment tool which allows researchers evaluating outcomes to control for variability of trauma severity (Senkowski and McKenney, 1999) and has an important role for comparative analyses used in quality control programs (Lefering, 2002).

The ISS is by far the most frequently applied score system for injury severity worldwide (Lefering, 2002), but it has been criticized and alternative scores proposed because it only assigns the worst abbreviated injury score to a severe single body area. Thus, a young adult with an abdominal gunshot wound causing injuries to the kidney, duodenum, vena cava, and pancreas would only have a predicted probability of mortality of 9%, provided all of the individual injuries were assigned non-fatal AIS codes.

Two scores to address that criticism have been proposed. The first is the Anatomic Profile (AP) (Copes et al., 1990) which is calculated using all serious injuries, defined as AIS > 2, to predict the probability of a patient’s survival. The AP discriminates survivors from non-survivors better than the ISS, but neither score alone reliably predicts patient outcomes. The second is the new injury sever-
ity score (NISS) (Osler et al., 1997) which uses the patient’s three most severe injuries, regardless of body region to predict the patient’s probability of mortality. It better predicts survival and is easier to calculate than the ISS.

A second criticism is the data for the AIS, and therefore the ISS, are not routinely collected by American hospitals (Rutledge, 1995). Data are costly because they are collected manually by coders who assign a set of injury codes to a case after reviewing information from medical records and other primary sources (Garthe et al., 1999). An experienced, trained coder can code three to four cases per hour (Garthe et al.). Training is a significant investment for any hospital. According to an advertisement on the Association for the Advancement of Automotive Medicine website, their ISS training course includes 14 hours of classroom work.

An alternative is to construct injury severity scores using diagnostic codes which are widely available in hospital computer systems. The first score to use diagnostic codes was the Anatomic Index (AI) (Champion et al., 1980b). It was derived from the single worst injury as measured by the Hospital International Classification of Disease Adaption (HICDA-8). The ICD-9 Injury Severity Score (ICISS) (Osler et al., 1996) is a modification of the AI that uses the Survival Risk Ratios (SSRs) derived from the International Classification of Disease, Ninth Revision, Clinical Modification (ICD-9-CM) codes for trauma patients to predict the probability of survival of injured patients. ICISS’s performance is superior to other trauma scores (Hannan et al., 2005; Expert Group on Injury Severity Measurement, 2005) and the score has been successfully adapted for the newer ICD-10 and ICD-10-AM diagnostic codes (Stephenson et al., 2004).

There are two concerns about ICISS. The first is with the SSRs which were calculated using a trauma registry of 314,402 patients with 504,934 injuries. Each SSR is the ratio of the number of survivors with the ICD-9-CM trauma code of interest to the number of patients in the database with that code (Osler et al., 1996). Thus, the calculation of the SSRs is database specific and possibly confounded with the outcomes of other injuries in multiple trauma cases (Expert Group on Injury Severity Measurement, 2005). This concern is partially addressed by Meredith et al. (2003) who derived SSRs from incidents where patients sustained an isolated injury only. The second concern is about the use of ICISS in quality improvement
programs. Iezzoni (1997) also found that, if risk models for patient mortality following admission to hospital were based on hospital discharge abstracts, they had better statistical performances than models derived from clinical records. However, she cautions that such models should not be used if the risk adjustment is part of a scheme to monitor performance because they adjust for conditions arising prior to and those arising during the hospital stay—possibly because of iatrogenic complications.

Models Using the TRISS Method

The Trauma Score and Injury Severity Score (TRISS) methodology (Boyd et al., 1987) uses anatomical, physiological, and age characteristics to quantify the probability of survival as related to severity of injury. The algorithm (Institute for Algorithmic Medicine, 2006; Société Française d’Anesthésie et de Réanimation, 2006) to calculate the probability $p_j$ of survival for trauma victim $j$ is given as

$$p_j = \beta_0 + \beta_1 RTS_j + \beta_2 ISS_j + \beta_3 Age_j$$

where $RTS_j$ is the Revised Trauma Score, $ISS_j$ is the Injury Severity Score, and $Age_j$ is a dichotomized variable with age $\leq 54$ years coded 0 and $\geq 55$ coded 1. Boyd et al. (1987) dichotomized at 55 because unpublished work showed that, for comparable levels of physiological derangement and anatomical injury severity, significantly increased mortality was associated with patients aged greater than 55.

There are two refinements. The first is that there are separate estimates of the parameters $\beta_i$, $i = 0, \ldots, 3$, for victims of blunt and penetrating trauma. The second is that the risk factors of the Revised Trauma Score for the TRISS model are weighted such that the expected probability of mortality of a patient with serious head injuries is more accurate than that predicted by the RTS for field triage.

The TRISS is proposed as a standard approach for evaluating outcome of trauma care (Boyd et al., 1987) quantified by physiological (RTS) and anatomical (ISS) characteristics plus age as a surrogate for physiological reserves. However, estimation of the TRISS model parameter values, called “norms”, has been an issue for its use in the evaluation of and quality assurance activities for trauma
care (Champion et al., 1990b). There have been many studies of the probabilities of mortality predicted by Equation (2.1).

The parameter values of Equation (2.1) were originally estimated using data from the Major Trauma Outcomes Study (MTOS). It is a retrospective descriptive study of injury severity of major trauma patients and is co-ordinated through the American College of Surgeons Committee on Trauma (Champion et al., 1990b). The parameter estimates obtained by Boyd et al. (1987), who used a dataset of over 25,000 cases submitted to MTOS in 1986, were \((-1.2470, 0.9544, -0.0768, -1.9052)\) for blunt trauma patients. In a second study using MTOS data, Champion et al. (1990b) used data for 80,544 trauma cases between 1982 and 1987 to estimate the TRISS norms as \((-1.3054, 0.9756, -0.0807, -1.9829)\). In a further study using MTOS data where the injury severity score was computed using AIS-90 rather than AIS-85 injury codes, Champion et al. (1995) estimated the norms as \((-0.4499, 0.9085, 0.0835, -1.7430)\). These results provide an illustration that, although the norms of the TRISS were estimated using large subsets of data from the MTOS database, estimation of the parameter values of logistic regression equations depends on the dataset used. That is, the probabilities of an adverse outcome predicted by the risk model are uncertain.

There are many other studies of the outcomes for blunt trauma victims (Hannan et al., 1995; Jones et al., 1995; Sugrue et al., 1996; Garber et al., 1996; Talwar et al., 1999; Onwudike et al., 2001; Joosse et al., 2001; Schall et al., 2002; Zafar et al., 2002; Podang et al., 2004; Murlidhar and Roy, 2004; Scarpelini et al., 2006; Skaga et al., 2006) that report the probabilities of survival of their trauma victims are different to those predicted by TRISS with norms derived from the MTOS database.

TRISS has been adopted by several local or national trauma registries and for comparison of different institutions (Lefering, 2002), but Hannan et al. (1995) concluded that the TRISS model is not sufficiently accurate to predict survival. They suggested that new variables be considered to improve the prediction for specific injury types, such as low falls.

Champion et al. (1990a) proposed A Severity Characterization of Trauma (ASCOT) as an improved anatomical and physiological scoring system that was more
complex than TRISS. In comparative studies with TRISS, Champion et al. (1996) found that ASCOT’s better calibration and more accurate predictions of patients’ probabilities of survival outweighed the greater complexity in computing the expected probabilities. They concluded that ASCOT should replace TRISS. On the other hand, Markle et al. (1992) compared both systems using data for trauma patients treated at hospitals affiliated with the New York Medical College. They concluded that ASCOT’s small gain in predictive accuracy is offset by its complexity and neither index provided good statistical agreement between predicted and observed outcomes for both blunt and penetrating injury patients.

From the perspective of statistical process control of processes to treat trauma patients, Markle et al. (1992) noted that those models were the best risk adjustment methods available and therefore useful for quality assurance, but concluded a new, more comprehensive index must be developed and tested on several datasets. Such an index would include additional variables, would adjust for the effects of several injury causes, and would incorporate decision rules for variable inclusion and weighting based on patient profiles. Markle et al. (1992) did not provide examples of additional variables. They did state that the difficulties of developing an index that was more comprehensive than ASCOT or TRISS cannot be underestimated. Hannan et al. (1995, 1997) considered a strategy of recalibrating a risk model to the population to be feasible, provided the recalibration used data collected after the process was in a state of statistical control. Hannan et al. (1997) suggested that, rather than search for a risk model that could be applied universally, regions develop statistical models tailored to their own patients. For example, they used their data to fit a logistic regression model that predicts mortality better than TRISS or ASCOT.

2.4 Conclusion

This chapter commenced by reviewing the adaption of control charts described in text books on statistical process control for industrial processes to the medical context. Such charts have been used to monitor clinical quality characteristics, such as surgical site infections and length of stay in a hospital, and administrative quality
characteristics, for example, ward occupancy. Because adverse medical events are, in general, quite rare, monitoring schemes, such as the EWMA control chart and the sets method, have been developed to monitor the number of normal events between the adverse events. The risk of adverse medical outcomes is known to vary between patients. Risk adjusted control charts, such as the VLAD chart and the RSPRT control chart, have been developed to control patient heterogeneity in monitoring schemes that use adverse medical outcomes as the quality characteristic. The review of control charts used in the medical context ended with descriptions of schemes, for example, funnel plots and profiling allowing for regression to the mean, proposed to control the between hospital variation and overdispersion associated with schemes used to compare performances of institutions or clinicians.

The second part of this literature review concerned the use of risk adjustment to control for the risk of adverse patient outcomes. So-called severity measurement systems are used to quantify the risks of the short-term outcomes for medical and surgical patients. This review provided examples of outcomes for which risk models had been constructed and gave a detailed assessment of risk scores used to quantify the risk of mortality of trauma victims. There has been no risk score for the triage of trauma patients that has been universally accepted. For example, the guidelines set by the American College of Surgeons for the transport of trauma patients to a major trauma centre are very conservative. They state that a rate of 50% for overtriage is acceptable. The trauma centres have responded by a process in which an experienced clinician, who is in contact with the paramedics as the patient is being transported to the trauma centre, continually assesses the severity of patient’s injuries and decides if the patient will require treatment in the ICU and/or operating theatre. The chapter concluded with a review of trauma scores used in quality improvement programs. They tend to be more complex than triage scores, but no current model predicts survival accurately and all need recalibration when used to adjust the risk of mortality of populations that did not contribute to the developmental databases used to construct the trauma scores.
Chapter 3

Using Meta-Factors to Compare Risk Models that Adjust for Outcomes of Cardiac Surgery

In Chapter 2, the literature review of the severity measurement systems for trauma victims was descriptive only. That review found 25 pre-hospital trauma scores, five risk indices modelled on the Injury Severity Score’s method of risk adjustment (Baker et al., 1974), and two models based on the TRISS method (Boyd et al., 1987). Terregino (1998) found that the proliferation of triage scores, in particular, was probably motivated by a continuing quest for a perfect score, but no model was found to be consistently better than the others or had been universally accepted.

This chapter continues the review of the risk-adjustment literature. It reviews the literature which describes the construction of various cardiac surgical risk models and assesses the performance two of those models—the Parsonnet score and the EuroSCORE. In contrast to the review of the trauma score literature in Chapter 2, the candidate undertakes an analysis of the information provided in publications on cardiac surgical risk scores to draw some conclusions about their performance and to propose a new method of risk factor selection for the construction of cardiac surgical risk models.
3.1 Introduction

In the late 1980s, concerns about unexplained variations in the use of expensive procedures and interventions across groups of patients led to calls for investigations into causes and consequences of varying health care practices in the United States (Iezzoni, 1994c). These concerns initiated the current era of assessment and accountability for American health care following eras of expansion in the late 1940s to 1960s and cost containment in the 1970s to the early 1980s. In the context of assessment and accountability, the emphasis is on the clinical outcomes and costs of health care. Iezzoni notes that the hallmark of this process is its concentration on the outcomes of services delivered by providers in the general community or in usual and customary practice settings—the so-called *effectiveness* of care. She contrasts this to the interest in the *efficacy* characteristic of controlled clinical trials, where treatments are administered by experts through tightly specified protocols to selected patients in closely monitored practice settings under trial conditions.

Iezzoni (1994c) states meaningful assessment of patients’ outcomes requires a measure of the *outcome* itself and a way to adjust for patients’ *risks* for the outcome of interest. She uses the example of the furore accompanying initial public release, in 1986, of hospital-level mortality by the Health Care Finance Administration (HCFA) to highlight the consequences of failing to account adequately for the variation in patients’ risks. According to the HCFA, 127 hospitals had significantly lower death rates, 142 had significantly higher rates, and, at the facility with the most aberrant rate, 87.6% of Medicare patients died compared to a rate of 22.5% predicted by the HCFA model. This facility was a hospice caring for terminally ill patients. Iezzoni states it is apparent inferences were made about the effectiveness of care using a model that did not adequately capture patients’ risks of death. The example she gives illustrates the importance of allowing for patients’ risks before comparing outcomes.

Risk adjustment models are constructed using patient risk factors appropriate for the outcome being measured to control for varying probabilities that patients will experience adverse outcomes following the procedure or intervention of interest. The assumption is that risk adjustment removes one source of variation in
the probabilities of outcomes and that the residual variation is due to the quality of care and random chance (Iezzoni, 1997). That is, if patients’ risks are adequately controlled, appropriate outcome measures are useful for the assessment of the quality of treatment or care within an institution.

Lilford et al. (2004) criticize the use of outcome data as a measure of quality of care. They note outcomes are influenced by definitions, data quality, patient case-mix, clinical quality of care, and chance and they state that, with the exception of chance, each of these sources of variation have some components that can be measured and others that cannot. They conclude that judgments about the quality of care on the basis of risk adjusted comparison cannot guarantee like is being compared with like because, even if a risk adjustment method could be agreed, it is possible the presence of unmeasured factors is causing systematic, between provider variation in measured outcomes. They state the sensitivity of the an institution’s position in league tables to the method of risk adjustment used suggests that comparisons of outcomes are unlikely to tell much about the quality of care.

Coronary artery disease has been known to be an important problem for a long time. In the United States, outcome measures have been used to assess the quality of care provided to cardiac surgical patients since the late 1980s when the treatment of coronary artery disease was identified as an important medical problem (Hannan et al., 1990). They identified two major quality of care issues for patients with cardiovascular disease. The first is the appropriateness of the chosen treatment to provide a long term benefit to the patient. For example, an indication of inappropriate cardiac surgical treatment is the significant differences in the rates of angiographies, angioplasties, and bypasses between black and white patients hospitalized in the United States with coronary artery disease (Hannan et al., 1990). They stated the issue of appropriate treatment has been studied in clinical trials to evaluate morbidity and mortality differences between patients undergoing coronary angioplasty and those undergoing coronary artery bypass grafts (CABGs) and, in 1990, research was being undertaken to develop appropriate criteria for CABG.

The second quality issue concerns the standard of care given. It is currently assessed in New York State by monitoring the short-term risk of mortality following
cardiac surgery (NY State, 1994-2004). Hannan et al. (1990) assessed the standard of care provided by New York State cardiac surgical centres by comparing their actual mortality rates with their predicted mortality rates, where the predicted probability of mortality of each patient was calculated from a logistic regression model with thirteen risk factors as covariates. The regression parameters were estimated using the data from the 28 New York State cardiac surgical centres in the study. The model was used in deriving risk adjusted mortality rates (defined and described in Chapter 4.2.1) from which four of the 28 hospitals were identified as “high outliers” and three as “low outliers”. Hannan et al. (1990) found that their model, which was constructed using risk factors derived from the patients’ clinical records, had predictive power superior to that of a model constructed using New York’s discharge data set.

There have been several statistical approaches to the retrospective or prospective use of pre-operative risk factors for the estimation of operative mortality in open heart surgery (Bernstein and Parsonnet, 2000). They note that predictive models are often used to compare the quality of care provided by institutions and surgeons in attempts to avoid the misunderstandings that can result from consideration of raw mortality rates alone. They identify four problems in modelling pre-operative risk. The first is the need for clearly defined and uniformly applied definitions for data collection. In an earlier publication, Higgins et al. (1992) stated there was a lack of standardized criteria for comparing outcomes in relation to preoperative conditions limits comparisons between institutions or different therapeutic approaches in patients undergoing coronary artery surgery. The second problem is the inclusion of covariates based or partly based on subjective judgment. Such risk factors contribute to uncertainty that may compromise the validity of the model. The third issue is collinearity if any explanatory variables are represented implicitly elsewhere in the model and the fourth is overestimation of pre-operative risk. Bernstein and Parsonnet (2000) state that overestimation of the probability of an adverse outcome is the result of including generalized or subjective risk factors that duplicate or reinforce other covariates already present in the model.

This chapter reviews several publications proposing models with pre-operative
patient characteristics to predict adverse outcomes following cardiac surgery. It also reviews articles investigating the performance of two risk scores in populations and times remote from the patient population(s) from which data used to construct the scores were collected. The first objective of this study is to identify the broad sets of patient characteristics that reflect the various sources, also known as dimensions (Iezzoni, 1994a), of risk for adverse outcomes after cardiac surgical procedures. The second is to propose a procedure using the dimensions identified to develop a risk adjustment score for a program of process improvement of cardiac surgery undertaken in a specific hospital. The premise is that customizing the risk adjustment for a particular institution will control for those unmeasurable sources of variation in outcome measures as described by Lilford et al. (2004).

The literature review is in three sections. In Section 3.2, in-depth descriptions of three models provide examples of the increase over time in the use of statistical methods for the construction of cardiac surgical risk scores. The first score described is the simple classification proposed by anaesthetists at the Montréal Heart Institute (Paiement et al., 1983), which classifies risk according to the number of risk factors present; the second is the Parsonnet score (Parsonnet et al., 1989), which pioneered systematic risk stratification (Gogbashian et al., 2004); and the third is the EuroSCORE (Nashef et al., 1999), which was developed as a robust scoring system for a wide population of cardiac surgical patients and is based on explicit objective criteria (Gogbashian et al., 2004).

Iezzoni (1994a) suggested there are broad sets of patient characteristics which reflect various dimensions or sources of risk. Depending on the outcome of interest, one or more of these dimensions of risk should be considered as indicators of patient risk. She states that evaluating the potential linkage between the different dimensions and the outcome is an essential step toward assessing the “medical meaningfulness” of the risk adjustment strategy. In Section 3.3, there is an assessment of the performance of 27 cardiac surgical risk scores for mortality. To assist in the assessment, the risk factors used in the scores were classified into eight categories, where the choice of categories was based on “medically meaningful” dimensions of risk for mortality following cardiac surgery. In Section 3.3.1, the proposed categories, which were called meta-factors, are defined. In Section 3.3.2
the covariates of 27 cardiac surgical risk scores are classified into their meta-factor categories to show that all of the scores were constructed using explanatory variables from similar broad sets of patient characteristics that contribute to the risk of adverse outcomes following cardiac surgery.

In Section 3.4, the candidate undertook a meta-analysis of the performances of the Parsonnet and EuroSCORE in patient populations remote from those that provided their training data are reviewed. If either model is used to risk adjust the mortality outcomes of cardiac surgical patients, changes to the risk profile of the patient population do not affect their effectiveness in ranking patients’ risks of mortality, although the measures of the predictive power for the EuroSCORE are generally a little higher than those for the Parsonnet score. However, if they are used to predict probabilities that cardiac surgical patients will die following their operation, both tend to overestimate probabilities of mortality in populations remote from those used for their construction, although the Parsonnet score generally gives higher predicted probabilities than the EuroSCORE.

The results from the meta-analyses undertaken in Sections 3.3 and 3.4 motivated the two proposals in Section 3.5. The first is that an institution which is using an “off-the-shelf” risk-adjustment score, such as the EuroSCORE, should recalibrate the score using data from cardiac surgical patients treated at that institution. The recalibrated score will provide mean expected probabilities of mortality which will be equivalent to the institution’s observed post-operative mortality rate. The alternative is to use the data collect by the institution to construct a customized risk adjustment score with a limited number of risk factors selected such that the risk factors in the risk model provide a measure for each of the “medically meaningful” dimensions of risk identified in Section 3.3.1. This chapter concludes with some discussion in Section 3.6.
3.2 Descriptions of Three Risk Models

3.2.1 A Simple Classification

A simple classification for the risk of in-hospital mortality was introduced by the anaesthetists at the Montréal Heart Institute in 1980 (Paiement et al., 1983). The score takes integer values from 0 to 8, where the integer value is given by the number out of eight risk factors, which are known to be associated with increased post-operative morbidity and mortality, present at the patient’s pre-operative assessment. A patient’s risk of in-hospital mortality is classified as normal, increased, or high for the respective scores of 0, 1, or 2 or more. Paiement et al. (1983) found in a sample of 500 patients that the respective probabilities of mortality for those with normal risk, those with increased risk, and those with high risk were 0.4%, 3.1%, and 12.2%.

In a follow up study of 2,029 patients in 1990, Tremblay et al. (1993) found that the proportion of high risk patients undergoing cardiac surgery had trebled and the proportion of normal risk patients had decreased by one third. They also found that the mean probability of mortality for normal patients had increased to 0.7%, but had decreased to 1.4% and 7.4% for patients with increased risk and high risk, respectively. They reported, as Paiement et al. (1983) did, that the risk of mortality for patients with a score of two or more was twelve times the risk for patients with a score of zero. Tremblay et al. (1993) concluded new and improved therapeutic modalities permitted better quality of care in 1990, but the Montréal Heart Institute risk assessment classification remained a useful clinical tool to predict hospital mortality and morbidity after cardiac surgery and to evaluate the quality of care within the institution.

In summary, the simple classification of cardiac surgical risk

- is a count of the number of the patient’s pre-operative risk factors;
- is constructed with medically meaningful risk factors;
- classifies cardiac surgical patients into three levels of risk of postoperative mortality; and
• shows comparable discrimination when used for populations with varying risk profiles.

3.2.2 The Parsonnet Score

The Parsonnet score (Parsonnet et al., 1989) was proposed as a relatively straightforward additive score that gives estimates of the probability that adult patients who undergo open heart surgery will die within 30 days of their operation. The sixteen risk factors for mortality following an operation given in Table 3.1 were chosen using a sample of 3,500 patients who underwent cardiac surgery between 1982 and 1987. The five selection criteria for choosing the risk factors were:

• their predictive value must be demonstrated by univariate analysis;

• they must be as free as possible from subjectivity or bias;

• they must be simple and direct (not derived from other information);

• risk factor data must be available at every hospital; and

• risk factor data must be available for every patient.

The weight for each risk factor in Table 3.1 is a univariate approximation of the odds ratio of patient death with the risk factor present relative to the risk factor absent. The Parsonnet score, \( p_t \) for patient \( t \), is found by summing the weights of the risk factors found in patient \( t \), is limited to values from 0 to 100, and gives the approximate percentile risk that patient \( t \) will die.

The additive score was tested against a multivariable logistic regression model (Parsonnet et al., 1989, Table 1) which was constructed using the same sample of 3,500 patients. It used 17 risk factors as explanatory variables. The differences between the two models are that the risk factor, other rare circumstances (e.g., paraplegia, etc.), is included in the additive model only, and the risk factors, ‘elevated cholesterol’ and ‘smoking’, are included in the multivariable model only. Parsonnet et al. (1989) tested both models using a data set containing 300 cases who had undergone cardiac surgery at an earlier date than those in the training
Table 3.1: Risk factors and weights for Parsonnet score.

Source: Parsonnet et al. (1989, Table 2)

This table is not available online. Please consult the hardcopy thesis available from the QUT Library.

continued on next page
data set. They found a correlation of 0.99 between the expected probabilities of mortality obtained from each of the models.

A second study using prospective data collected from 1,332 patients undergoing open heart procedures at the Newark Beth Israel Hospital showed that the observed death rates and the probabilities of mortality predicted by the Parsonnet score were highly correlated (Parsonnet et al., 1989, Figure 3). However, the authors noted that the actual results of surgery were better than predicted. They concluded that their original weights were incorrect or the operative results of the institution had improved “over the past two years”. They said that it is possible that periodic correction in the weighted factors will be necessary.

Parsonnet et al. (1989) proposed the additive risk model given in Table 3.1 because it was of practical use and could be easily applied to any hospital or surgeon. Its purpose is to stratify a heart surgery patient preoperatively into one of five risk groups, where the ranges of Parsonnet scores 0–4, 5–9, 10–14, 15–19, and ≥20 were chosen to classify the risk of postoperative mortality into levels of good, fair, poor, high, and very high, respectively. There were no reasons given for the choice of the range of Parsonnet scores for each category of risk. Parsonnet et al. found there was a satisfactory similarity of results at the three hospitals that participated in the studies. They concluded that comparisons between individual surgeons were feasible.

In summary, the Parsonnet score

- gives the pre-operative percentile probabilities of postoperative mortality of
cardiac surgical patients as the sum of integer weights assigned to each of
the risk factors present at the patients’ pre-operative assessment;

• has five criteria, including univariate empirical analysis, for selection of risk
factors;

• estimates the weight of each risk factor as an integer approximation of the
odds ratio for mortality in the presence of the risk factor relative to the
absence of the risk factor;

• classifies patients into five levels of pre-operative risk; and

• showed comparable performance at three institutions.

3.2.3 The EuroSCORE

The European System for Cardiac Operative Risk Evaluation (EuroSCORE) is a
risk stratification system for the expected mortality of all cardiac surgical patients.
It was constructed as a tool for the assessment of the quality of cardiac surgical
care (Nashef et al., 1999). Roques et al. (1999) stated that they embarked on the
study to establish the risk profile of adult European cardiac surgical patients and
determine the procedural mortality of those patients. In their opinion, little was
known about the risk profile of European cardiac surgical patients.

In the first stage of its development, Roques et al. (1999) used a data set
with 97 potential explanatory variables and 19,030 patients, who underwent car-
diac surgery at 128 centres in eight European countries between September and
December 1995, to fit a logistic regression model with patient mortality as the
response variable. Comprehensive information on data collection and definitions
of variables was provided to all participating institutions and the database was
audited to confirm that the data collected were complete and > 99% accurate.
Roques et al. (1999, Appendices B–H) provide definitions of their 97 explanatory
variables and, in Appendix I, they define mortality as death prior to the patient’s
discharge from hospital or death within 30 days of the operation if the patient has
already been discharged. They constructed a logistic regression model with nine-
teen risk factors—the seventeen given in Table 3.2 plus chronic congestive cardiac
Table 3.2: Risk factors, risk factor descriptions, risk factor weights (Wt) as given by Nashef et al. (1999, Table 2) and parameter estimates ($\beta$-Coef) as given by Michel et al. (2003, Table 1) for the EuroSCORE. Note that the $\beta$-coefficients were given to 7 significant figures.

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Description</th>
<th>Wt</th>
<th>$\beta$-Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patient-related factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Active Endocarditis</td>
<td>Patient still under antibiotic treatment for endocarditis at time of surgery</td>
<td>3</td>
<td>1.101265</td>
</tr>
<tr>
<td>Age</td>
<td>Per 5 years or part thereof over 60 years</td>
<td>1</td>
<td>0.0666354</td>
</tr>
<tr>
<td>Chronic pulmonary disease</td>
<td>Longterm use of bronchodilators or steroids for lung disease</td>
<td>1</td>
<td>0.4931341</td>
</tr>
<tr>
<td>Critical preoperative state</td>
<td>Any one or more of the following: ventricular tachycardia or fibrillation or aborted sudden death, preoperative cardiac massage, preoperative ventilation before arrival in the anaesthetic room, preoperative inotropic support, intraaortic balloon counter pulsation or preoperative acute renal failure (anuria or oliguria &lt; 10ml/h)</td>
<td>3</td>
<td>0.9058132</td>
</tr>
<tr>
<td>Extracardiac arteriopathy</td>
<td>Any one or more of the following: claudication, carotid occlusion or &gt; 50% stenosis, previous or planned intervention on the abdominal aorta, limb arteries or carotids</td>
<td>2</td>
<td>0.6558917</td>
</tr>
<tr>
<td>Neurological dysfunction</td>
<td>Disease severely affecting ambulation or day-to-day functioning</td>
<td>2</td>
<td>0.841626</td>
</tr>
<tr>
<td>Previous cardiac surgery</td>
<td>Requiring opening of the pericardium</td>
<td>3</td>
<td>1.002625</td>
</tr>
<tr>
<td>Serum Creatinine</td>
<td>$&gt; 200\mu$mol/l preoperatively</td>
<td>2</td>
<td>0.6521653</td>
</tr>
<tr>
<td>Sex</td>
<td>Female</td>
<td>1</td>
<td>0.3304052</td>
</tr>
</tbody>
</table>

continued on next page
**Table 3.2:** Risk factors, risk factor descriptions, risk factor weights (Wt), and parameter estimates ($\beta$-Coef) for the EuroSCORE (continued).

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>Description</th>
<th>Wt</th>
<th>$\beta$-Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cardiac-related factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LV Dysfunction</td>
<td>Moderate or LVEF 30–50%</td>
<td>1</td>
<td>0.4191643</td>
</tr>
<tr>
<td></td>
<td>Poor or LVEF &lt; 30%</td>
<td>3</td>
<td>1.094443</td>
</tr>
<tr>
<td>Pulmonary hypertension</td>
<td>Systolic PA pressure &gt; 60mm Hg</td>
<td>2</td>
<td>0.7676924</td>
</tr>
<tr>
<td>Recent myocardial infarction</td>
<td>&lt; 90 days</td>
<td>2</td>
<td>0.5460218</td>
</tr>
<tr>
<td>Unstable angina</td>
<td>Rest angina requiring i.v. nitrates until arrival in anaesthetic room</td>
<td>2</td>
<td>0.5677075</td>
</tr>
<tr>
<td><strong>Operations-related factors</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency</td>
<td>Carried out on referral before the next working day</td>
<td>2</td>
<td>0.7127953</td>
</tr>
<tr>
<td>Other than isolated CABG</td>
<td>Major cardiac procedure other than or in addition to CABG</td>
<td>2</td>
<td>0.5420364</td>
</tr>
<tr>
<td>Postinfarct septal rupture</td>
<td></td>
<td>4</td>
<td>1.462009</td>
</tr>
<tr>
<td>Thoracic aortic surgery</td>
<td>For disorder of ascending or descending aorta</td>
<td>3</td>
<td>1.159787</td>
</tr>
</tbody>
</table>

failure and urgent operation—for postoperative mortality. They state the covariates that they entered into the model were chosen using bivariate tests, chi square tests for categorical variables, and unpaired tests for continuous covariates. All variables significant at the $P < 0.2$ level were entered into the model provided they were present in at least 2% of the sample, then non-significant variables were removed using a process of backward elimination. There was no allowance in the model for between hospital or between country effects.

In the second stage, an international panel of cardiac surgeons with an interest in risk stratification evaluated the selected risk factors using four criteria—objectivity, credibility, availability, and resistance to falsification (Nashef et al., 1999). The definitions used for these criteria were not given, but the congestive
cardiac failure and urgent operation variables were discarded because they were deemed liable to distortion. The weights of the additive EuroSCORE given in Table 3.1 were attributed using the $\beta$-coefficients of a multivariable logistic model fitted to a developmental, or training, data set of 13,302 cardiac surgical patients and they were validated using a test data set of 1,479 patients. The precise method used to assign the weights and a detailed description of the data set are not given. It is clear, however, that the training and test data are from those patients who underwent operations at cardiac surgical centres in Germany, France, Italy, Spain, and Finland between September and December, 1995. Patients with EuroSCOREs of 0–2, 3–5, and $\geq 6$ are classified as low, medium, and high risk, respectively.

Nashef et al. (1999) state that the EuroSCORE is a simple and objective system with risk factors that are measurable, readily available, and resistant to falsification because all except four are derived from the clinical status of the patient. The other four factors are related to the operation, but are difficult to influence through subtle variation in surgical decision making. Consequently, they recommended the EuroSCORE as a suitable tool for monitoring the quality of cardiac surgical practice. Because they understood the need to assess the performance of the EuroSCORE in a variety of patient populations, Nashef et al. (1999) invited other workers to test it in their hospitals and on individual patient and procedural subgroups.

Michel et al. (2003) note that, although the additive EuroSCORE is a simple and accessible-to-all “gold standard” of risk assessment, it has been found to underestimate risk in certain very high-risk patients. They identified the additive nature of the standard EuroSCORE as the reason for the underestimation and proposed the logistic EuroSCORE for predicted probabilities of mortality greater than 8 to 10%. For the logistic EuroSCORE, the predicted probability of mortality $p_t$ is given by

$$p_t = \logit^{-1} \left( \alpha + \sum_{i=1}^{17} \beta_i X_{i,t} \right),$$

where $X_{i,t}$ are the risk factors for patient $t$ and $\alpha$ and $\beta_i$ are the regression coefficients estimated from the original EuroSCORE data to seven significant figures. The estimate for the intercept is $\alpha = -4.78954$ and the estimates of the risk
factor parameters, $\beta_i$, are given in Table 3.2. Use of the logistic EuroSCORE is recommended for calculation of realistic probabilities of mortality for very high risk patients, for monitoring and risk stratification where high-risk surgery forms a substantial part of the work load, and for research purposes.

In summary, the additive EuroSCORE

- gives the pre-operative percentile probabilities of postoperative mortality of cardiac surgical patients as the sum of integer weights assigned to the patients’ risk factors;

- uses multivariable logistic regression followed by assessment by an expert panel for selection of risk factors;

- approximates the weights from the multivariable odds ratios;

- classifies patients into three levels of pre-operative risk; and

- was proposed as a suitable tool for monitoring cardiac surgical practices.

### 3.3 Classification of Cardiac Surgical Risk Factors

A literature review of pre-operative risk adjustment for cardiac surgery (Djamaludin, 2006) identified 31 risk models; all of which use patient characteristics, only, to predict the risk that patients will experience adverse postoperative outcomes. Reasons given for the construction of new cardiac surgical risk scores include:

- an accurate means for risk assessment has not been established (Edwards et al., 1988);

- scores devised in the past are too complicated for practical use (Parsonnet et al., 1989);

- current methods do not adequately explain if the differences between institutions are due to variations in patient severity or quality of care (Higgins et al., 1992);
• comparison of models constructed using logistic regression with those formulated using alternative methods (Marshall et al., 1994b);

• development of a simple index for inter-institutional comparisons of multiple outcomes (mortality and Length of Stay (LOS)) (Tu et al., 1995);

• updates and revisions of models to reflect the current clinical milieu (Edwards et al., 1997); and

• establishment of the current risk profile of European cardiac surgical patients (Roques et al., 1999).


The aim of this section is to compare the performances of the risk models in the literature review by Djamaludin (2006). That review also identified 181 risk factors for an adverse event following cardiac surgery. Because each risk factor had been used as a covariate in at least one of the 31 risk models in the literature review, fair comparison and assessment of their performances appears to be difficult. To facilitate a comparison of the performance of cardiac surgical models with different risk factors as explanatory variables, a hypothesis is proposed that each risk factor be classified into one of eight meta-factor categories for cardiac surgical risk. Prototype definitions for the meta-factors are given in Section 3.3.1. In Section 3.3.2, the performance of the risk scores is assessed. The number of meta-factors used to construct each risk model and, if published, the measure of its predictive power are used as assessment tools. The results from these two as-
assessments show that the construction and predictive power of the cardiac surgical risk scores in the literature review are comparable.

3.3.1 Description of Meta-Factors

In the discussion of the dimensions of risk, Iezzoni (1994a, Table 2.1) proposed that patient risk factors be classified into eleven sources of risk for health care outcomes. She notes that the inclusion or exclusion of any dimension in a risk model depends on an evaluation of its potential linkage with the particular outcome of interest and that such an evaluation is an essential step towards assessing the medical meaningfulness, or clinical credibility, of a risk-adjustment strategy. Graham and Cook (2004) note that the common underlying method of modelling risk of mortality in an intensive care unit (ICU) is to group and preprocess observations into their sources of risk. They identify acute physiological disturbance as the dimension of risk which provides most of the predictive power of risk models for ICU mortality and note that it may be measured by summing weighted scores of deviations from the normal physiological ranges, where the weights may be taken as those given for some published ICU risk score for mortality, such as the APACHE III score (Knaus et al., 1991). Other dimensions of risk for ICU mortality that may be included in the model are chronological age and measures of chronic or comorbid conditions, both of which control for the patients’ physiological reserves, plus diagnosis, surgical status, lead-time, and individual hospital characteristics.

When the EuroSCORE was proposed, eight of its risk factors, listed in Table 3.2, were categorized as patient-related factors, four as cardiac-related, and four as operations-related, but no publications on the classification of risk factors for cardiac-surgical risk-adjustment into sources of risk were found during the literature review. The candidate worked on the classification of all cardiac surgical risk factors into meta-factor categories with Dr David A. Cook, an intensive care specialist at the Princess Alexandra Hospital. He proposed the seven meta-factor classes and one miscellaneous class in Table 3.3 and categorized each of the 181 risk factors in Appendix A into its meta-factor class.

The proposed meta-factors are medically meaningful because they are measures
Table 3.3: Proposed meta-factor categories for cardiac surgical outcomes.

<table>
<thead>
<tr>
<th>Code</th>
<th>Meta-factor</th>
<th>Patient State</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Age</td>
<td>Physiological Reserve</td>
</tr>
<tr>
<td>2</td>
<td>Comorbidities</td>
<td>Physiological Reserve</td>
</tr>
<tr>
<td>3</td>
<td>Reduced Blood Supply to the Heart</td>
<td>Cardiac Reserve</td>
</tr>
<tr>
<td>4</td>
<td>Cardiac Failure/LV* Systolic Dysfunction</td>
<td>Cardiac Reserve</td>
</tr>
<tr>
<td>5</td>
<td>Other Complications Limiting Heart Function</td>
<td>Cardiac Reserve</td>
</tr>
<tr>
<td>6</td>
<td>Procedure/Diagnosis Factors</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>7</td>
<td>Acute Illness</td>
<td>Acute Illness</td>
</tr>
<tr>
<td>8</td>
<td>Unclassified</td>
<td></td>
</tr>
</tbody>
</table>

*Left Ventricular

which provide information about particular patient states—physiological reserve, cardiac reserve, diagnosis, and acute illness—that are relevant to the surgeons’ assessment of the risk of adverse outcomes following cardiac surgery.

When Dr Cook provided the classifications of the risk factors into their meta-factors, he suggested further research to ascertain the opinion of cardiac surgeons or clinicians who treat cardiac surgical patients as to appropriate meta-factors for adverse outcomes following cardiac surgery.

3.3.2 Performance of Cardiac Surgical Risk Scores

The literature review (Djamaludin, 2006) identified 23 studies that proposed new or modified existing scores for the risk of mortality following cardiac surgery. Fourteen of the studies were carried out in North America, eight in Europe, and one in Australia. The summary in Appendix B gives the abbreviated name of each risk score, the study citation, the response definition used in each study, and the number of institutions and number of patients contributing to the data used to construct
Table 3.4: Summary for the 27 cardiac surgical risk scores listed in Appendix B. Abbreviated names of the risk scores are listed under Score Names, the number of institutions contributing to the each score’s training dataset under Data Sources, and the value of the area under the receiver operating characteristic curve for each score under AROC. The number of risk factors in each meta-factor class, as defined in Table 3.3, and the total number of risk factors for each score are also listed.

<table>
<thead>
<tr>
<th>Score Name</th>
<th>Data Sources</th>
<th>AROC</th>
<th>No of Risk Factors By Meta-Factor Code</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMPLE</td>
<td>1</td>
<td>0.81</td>
<td>1 1 1 2 0 2 1 0</td>
<td>8</td>
</tr>
<tr>
<td>BAYES</td>
<td>1</td>
<td>1</td>
<td>1 1 3 2 1 5 4 1</td>
<td>18</td>
</tr>
<tr>
<td>PARS</td>
<td>3</td>
<td>1</td>
<td>1 4 0 1 2 5 2 1</td>
<td>16</td>
</tr>
<tr>
<td>SYS97</td>
<td>10</td>
<td>0.81</td>
<td>1 14 1 4 2 7 4 3</td>
<td>36</td>
</tr>
<tr>
<td>NYSTATE:90</td>
<td>28</td>
<td></td>
<td>1 2 3 2 0 3 0 2</td>
<td>13</td>
</tr>
<tr>
<td>NYSTATE:94</td>
<td>30</td>
<td>0.79</td>
<td>1 4 3 2 0 1 1 2</td>
<td>14</td>
</tr>
<tr>
<td>CSSS</td>
<td>&gt; 1</td>
<td>0.74</td>
<td>1 5 0 1 0 5 1 0</td>
<td>13</td>
</tr>
<tr>
<td>O’CONNOR</td>
<td>5</td>
<td>0.74</td>
<td>1 2 2 2 0 2 0 1</td>
<td>10</td>
</tr>
<tr>
<td>SCRSS</td>
<td></td>
<td></td>
<td>1 2 1 2 0 4 0 2</td>
<td>12</td>
</tr>
<tr>
<td>B-LOGIT</td>
<td>&gt; 1</td>
<td>0.72</td>
<td>1 3 4 6 1 3 1 0</td>
<td>19</td>
</tr>
<tr>
<td>ONTARIO</td>
<td>9</td>
<td>0.75</td>
<td>1 0 1 1 0 2 0 1</td>
<td>6</td>
</tr>
<tr>
<td>ONTARIO:N</td>
<td>16</td>
<td>0.78</td>
<td>1 2 2 1 0 2 0 1</td>
<td>9</td>
</tr>
<tr>
<td>CRS:MAGOV</td>
<td>1</td>
<td>0.86</td>
<td>1 9 0 4 1 2 2 1</td>
<td>20</td>
</tr>
<tr>
<td>STS:90</td>
<td>&gt; 1</td>
<td>0.78</td>
<td>1 4 3 4 1 4 0 2</td>
<td>19</td>
</tr>
<tr>
<td>STS:91</td>
<td>&gt; 1</td>
<td>0.78</td>
<td>1 4 4 4 0 3 1 2</td>
<td>19</td>
</tr>
<tr>
<td>STS:92</td>
<td>&gt; 1</td>
<td>0.78</td>
<td>1 7 3 4 1 6 3 1</td>
<td>26</td>
</tr>
<tr>
<td>STS:93</td>
<td>&gt; 1</td>
<td>0.78</td>
<td>1 7 4 4 1 6 1 2</td>
<td>26</td>
</tr>
</tbody>
</table>

*continued on next page*
**Table 3.4**: Summary for the 27 cardiac surgical risk scores listed in Appendix B (continued).

<table>
<thead>
<tr>
<th>Score Name</th>
<th>Data Sources</th>
<th>AROC</th>
<th>No of Risk Factors</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>STS:94</td>
<td>&gt; 1</td>
<td>0.80</td>
<td>1 10 4 5 0 5 2 2</td>
<td>29</td>
</tr>
<tr>
<td>European (Including UK) Risk Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PARS:MOD</td>
<td>42</td>
<td>0.70</td>
<td>1 16 3 3 6 10 4 3</td>
<td>46</td>
</tr>
<tr>
<td>FRENCH</td>
<td>42</td>
<td>0.74</td>
<td>1 1 1 1 2 7 1 0</td>
<td>14</td>
</tr>
<tr>
<td>EURO</td>
<td>128</td>
<td>0.79</td>
<td>1 4 2 1 1 4 2 2</td>
<td>17</td>
</tr>
<tr>
<td>PSP</td>
<td>1</td>
<td>0.79</td>
<td>1 7 6 2 1 6 1 1</td>
<td>25</td>
</tr>
<tr>
<td>CATAL</td>
<td>7</td>
<td>1 1 2 1 2 1 6 1 0</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>CRS:STAAT</td>
<td>1</td>
<td>0.77</td>
<td>0 2 0 1 1 1 0 1</td>
<td>6</td>
</tr>
<tr>
<td>NWENG</td>
<td>4</td>
<td>0.73</td>
<td>1 2 0 1 0 4 2 0</td>
<td>10</td>
</tr>
<tr>
<td>AMPHIA</td>
<td>1</td>
<td>0.84</td>
<td>1 1 1 1 0 3 0 1</td>
<td>8</td>
</tr>
<tr>
<td>ASCTS</td>
<td>5</td>
<td>0.79</td>
<td>1 3 0 0 0 2 0 1</td>
<td>7</td>
</tr>
</tbody>
</table>

There was some variation in the methods used to construct the risk scores. Table 3.4 shows that risk models were developed using data from a single institution in six studies, multiple institutions in sixteen studies, and an unknown number of institutions in the study by Tuman et al. (1992). Perusal of the list in Appendix B shows the data were collected from the population of patients who underwent CABG, only, in ten of the studies. In the other thirteen, they were collected from a population of adult patients undergoing other cardio-thoracic surgical procedures, such as repair or replacement of one or more valves, combined CABG and valve operations, or some other cardiac and/or thoracic procedure. There were also important differences in the definition of the mortality outcome. In ten of the studies death was defined as in-hospital, in four it was 30 days post-operation, and in three it was a combination of in-hospital and 30 days post-operation. That is,
in 17 of the studies patient mortality is defined in one of three ways. For three of the other six studies the outcome is broadly defined as operative mortality.

Despite the differences between the studies, perusal of Table 3.4 shows similarities in meta-factors used to construct the risk scores. All the risk scores reviewed except the Clinical Risk Score (Staat et al., 1999) included age as a meta-factor and all except the Ontario score (Tu et al., 1995) have at least one comorbidity meta-factor. That is, all the risk scores in the review have at least one measure of the patient’s physiological reserve as an explanatory variable. All scores except the model proposed by Reid et al. (2001) have some measure of the patient’s cardiac reserve. There is only one score that does not have cardiac failure/LV systolic function as a meta-factor, six scores that do not have reduced blood supply to the heart as a meta-factor and thirteen that do not have other complications limiting heart function as a meta-factor. There is at least one procedure/diagnosis meta-factor in all of the risk scores reviewed, but acute illness is a meta-factor in only thirteen of the 23 risk scores reviewed. It appears that measures of physiological reserve, cardiac reserve, and diagnosis are important to the assessment of the risk of cardiac surgical mortality, and that a measure of the acute illness state is less important.

The Area under the Receiver Operating Characteristic curve (AROC) is a summary measure of the predictive power of a model. Agresti (2002) states that the AROC is identical to another measure of predictive power, the concordance ($C$) index. Consider all pairs of observations $(i, j)$ such that $y_i = 1$ and $y_j = 0$ and the estimates $\Pr(Y_i = 1) = \hat{\pi}_i$ and $\Pr(Y_j = 1) = \hat{\pi}_j$. The $C$ index estimates the probability that the predictions and outcomes are concordant, that is, the observations with the larger $y$ also have the larger $\hat{\pi}$. A value of $c = 0.5$ means the predictions are no better than random guessing.

The AROC or its equivalent, such as the $C$ index, was published for 20 of the 27 models listed in Table 3.4. It ranged from a minimum of 0.70 for the modified Parsonnet score (Gabrielle et al., 1997) to a maximum of 0.86 for the Clinical Risk Score (Magovern et al., 1996). The data used to construct the modified Parsonnet score collected from a patient population undergoing adult cardiac surgery at one of 42 institutions and the data for the Clinical Risk Score were gathered from a
patient population undergoing CABGs at one hospital. For the 20 risk scores, the average of the AROCs was 0.77, the median AROC is 0.78, and the interquartile range is from 0.74 to 0.79.

Figure 3.1 shows that the AROC values published for the models in this literature review are not correlated with the number of risk factors in the models. They are scattered randomly about the median AROC value of 0.78. This plot emphasizes that the score with the lowest AROC, the modified Parsonnet score, used the most risk factors in its construction and that two the scores with the highest AROCs, the Magovern and Amphia scores, were constructed using data from only one hospital. The lower explanatory power of models constructed using data from multiple sources indicates that the data were more heterogeneous than those from a single institution and, therefore, the risk model could explain less of the variability in the data. It appears that any between-hospital effect in the data is a more important factor for modifying the explanatory power of risk adjustment models than the number of risk factors in the model.

The Society for Thoracic Surgeons (STS) published five versions of their risk assessment model for the years 1990–94 (Edwards et al., 1997). Their stated reason for the annual update of their cardiac surgical risk score was to accurately model their current data. Edwards et al. (1997) concluded that the then most recent model developed from the STS database provided a reliable and statistically valid method to stratify CABG patients into appropriate risk categories.

The number of patients in the STS data sets increased by a factor of five from about 24,000 in 1990 to approximately 118,000 in 1994. Each year’s data was used to construct a model using a stepwise selection procedure in which risk factors were entered and retained in the model if the Wald \( \chi^2 \) statistic had a \( p \)-value such that \( p < 0.2 \) and \( p < 0.1 \), respectively (Edwards et al., 1997). This model selection procedure and the increasing number of patients in the training data sets allowed the construction of models where, according to the Wald criterion, the number of significant risk factors increased from 19 in 1990 to 29 in 1994. However, Table 3.4 shows that the predictive power of the STS risk scores remained stable at 0.78 for the 1990 to 1993 models and increased marginally to 0.80 for the 1994 version. The predictive power of the STS risk score did not increase as more risk factors
Figure 3.1: Each of the AROC values given for the risk models in Table 3.4 plotted against the number of risk factors used to construct the risk score. The line AROC = 0.78 gives the median for the 20 AROCs plotted.

Note: (a) the two highest AROCs are for two of the three risk models, the Magovern and Amphia scores, constructed using data from one hospital, only, and
(b) the lowest is for the modified Parsonnet score constructed using data from 42 cardiac surgical teams.
CHAPTER 3: META-FACTORS

were added.

Furthermore, Table 3.4 shows that all five versions of the STS risk score had a risk factor for the meta-factor class Age, several for Comorbidities, several for Reduced Blood Supply to the Heart, several for Cardiac Failure/LV Systolic Dysfunction, and several for Diagnosis. Four of the five models had at least one risk factor for Acute Illness. Given the proposition outlined in Section 3.3.1 that each risk factor provides a measure of some “clinically meaningful” dimension of risk, one would not expect the addition of risk factors for the same meta-factor classes to increase the discrimination ability of a risk adjustment model.

In conclusion, this exploratory meta-analysis indicates that the predictive power of cardiac surgical risk models which have multiple risk factors from each meta-factor class—for example, the Society Thoracic Surgeons’ models—is no better that that of risk scores that only have one or two risk factors from each meta-factor category—for example, the Ontario score. In this analysis, the two scores with the best predictive power were the Magovern and Amphia scores. The AROCs of both these score were found using data form patients treated a single institution. The AROCs of all but one of the other risk scores in the review were found using data gathered from multiple institutions. This suggests that between institution variability has greater influence on the predictive power of risk scores than does the number of risk factors included in risk models.

3.4 Performance Review of Two Risk Scores at Multiple Locations

This section contains a review of the literature on the performance of the Parsonnet score and EuroSCORE in risk adjusting the expected mortality of cardiac surgical patients in populations remote from the population used for their derivation. A search of the literature revealed 21 studies that contained an assessment of the performance of the Parsonnet score, the EuroSCORE, or both. The Parsonnet score is evaluated in fourteen and the EuroSCORE in twelve. Appendix C lists the following information about each study in the review—(a) the citation and
an abbreviated name for the study, (b) the score(s) evaluated, (c) the mortality
definition used in the study, (d) the patient population which provided the data
for the study, and (e) the number of patients in the study data set and the data
collection period.

The purpose of the review in Section 3.4.1 is to evaluate the predictive power of
the Parsonnet score and to compare its predicted probabilities of mortalities with
the observed mortality rates in the study populations. A similar assessment of the
EuroSCORE is undertaken in Section 3.4.2.

3.4.1 Parsonnet Score

As noted in Section 3.2.2, Parsonnet et al. (1989) defined mortality as any death
occurring within 30 days of surgery, but that definition is used in only three of the
fourteen studies. For the other studies, mortality is defined as in-hospital in six, it
is defined as in-hospital or within 30 days of surgery if the patient was previously
discharged in four, and it is not clearly defined in the article by Pliam et al. (1997).

The case mix of the populations of cardiac surgical patients, from which data
were collected, varied. For four of the studies, data were collected from patients
undergoing CABG surgery only. For the other ten, it was collected from populations
undergoing valve surgery only, combined valve and CABG surgery, or other
cardiac surgical procedures. The number of cardiac-surgical institutions particip-
ating in each study varied from one hospital in seven studies to 42 cardiac surgery
teams in the study by Gabrielle et al. (1997), in which the data were used for evalu-
ating the Parsonnet score as well as constructing the modified Parsonnet score,
as described in Section 3.3.2.

Table 3.5 gives the AROC, the mean expected probabilities of mortality, where
the mean is the average of predicted probabilities of mortality for all patients
in a study, and the observed mortality rate from each the studies analysing the
performance of the Parsonnet score. There are nineteen items listed in the table
because Pliam et al. (1997) provide one analysis for a cardiac surgical population
undergoing a mixture of operations and another for patients undergoing isolated
CABG surgery and Peterson et al. (2000) provide the average of the observed and
Table 3.5: Number of data sources (#Centres), Area under Receiver Operating Characteristic Curve (AROC), data set mean of predicted probabilities of mortality (Expected) and observed mortality rate (Observed), and standardized mortality rate (SMR) for assessing the Parsonnet score.

<table>
<thead>
<tr>
<th>Study</th>
<th>Surgery</th>
<th>#Centres</th>
<th>AROC</th>
<th>Expected</th>
<th>Observed</th>
<th>SMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORR:95</td>
<td>CABG</td>
<td>2</td>
<td>0.74</td>
<td>0.092</td>
<td>0.037</td>
<td>0.402</td>
</tr>
<tr>
<td>GAB:97</td>
<td>CABG &amp; Valve</td>
<td>42</td>
<td>0.65</td>
<td></td>
<td></td>
<td>0.060</td>
</tr>
<tr>
<td>PLI:97/1</td>
<td>CABG &amp; Valve</td>
<td>1</td>
<td>0.80</td>
<td>0.143</td>
<td>0.040</td>
<td>0.279</td>
</tr>
<tr>
<td>PLI:97/2</td>
<td>CABG</td>
<td>1</td>
<td>0.80</td>
<td>0.134</td>
<td>0.037</td>
<td>0.276</td>
</tr>
<tr>
<td>WEI:97</td>
<td>CABG &amp; Valve</td>
<td>1</td>
<td>0.70</td>
<td></td>
<td></td>
<td>0.035</td>
</tr>
<tr>
<td>BRI:98</td>
<td>CABG</td>
<td>2</td>
<td>0.73</td>
<td>0.046</td>
<td>0.037</td>
<td>0.804</td>
</tr>
<tr>
<td>MAR:99</td>
<td>Cardiac Surgery</td>
<td>1</td>
<td>0.86</td>
<td>0.056</td>
<td>0.055</td>
<td>0.982</td>
</tr>
<tr>
<td>GEI:00</td>
<td>Bypass Surgery</td>
<td>1</td>
<td>0.76</td>
<td>0.102</td>
<td>0.040</td>
<td>0.392</td>
</tr>
<tr>
<td>PET:00</td>
<td>CABG G*</td>
<td>28</td>
<td>0.72</td>
<td>0.030</td>
<td>0.010</td>
<td>0.333</td>
</tr>
<tr>
<td>PET:00</td>
<td>CABG F†</td>
<td>28</td>
<td>0.72</td>
<td>0.045</td>
<td>0.030</td>
<td>0.667</td>
</tr>
<tr>
<td>PET:00</td>
<td>CABG P‡</td>
<td>28</td>
<td>0.72</td>
<td>0.060</td>
<td>0.045</td>
<td>0.750</td>
</tr>
<tr>
<td>PET:00</td>
<td>CABG H§</td>
<td>28</td>
<td>0.72</td>
<td>0.087</td>
<td>0.065</td>
<td>0.747</td>
</tr>
<tr>
<td>PET:00</td>
<td>CABG VH¶</td>
<td>28</td>
<td>0.72</td>
<td>0.230</td>
<td>0.130</td>
<td>0.565</td>
</tr>
<tr>
<td>WYN:00</td>
<td>Cardiac Surgery</td>
<td>4</td>
<td>0.74</td>
<td>0.069</td>
<td>0.035</td>
<td>0.507</td>
</tr>
<tr>
<td>KAW:01</td>
<td>Bypass Surgery</td>
<td></td>
<td>0.72</td>
<td>0.104</td>
<td>0.045</td>
<td>0.433</td>
</tr>
<tr>
<td>NAS:01</td>
<td>Open Heart Surgery</td>
<td>1</td>
<td>0.74</td>
<td>0.102</td>
<td>0.042</td>
<td>0.412</td>
</tr>
<tr>
<td>ASA:03/1</td>
<td>CABG</td>
<td>2</td>
<td>0.73</td>
<td>0.058</td>
<td>0.034</td>
<td>0.586</td>
</tr>
<tr>
<td>VAN:03</td>
<td>Cardiac Surgery</td>
<td>1</td>
<td>0.90</td>
<td>0.149</td>
<td>0.056</td>
<td>0.376</td>
</tr>
<tr>
<td>NIL:06</td>
<td>Cardiac Surgery</td>
<td>1</td>
<td>0.76</td>
<td>0.109</td>
<td>0.029</td>
<td>0.266</td>
</tr>
</tbody>
</table>

*Good Risk  †Fair Risk  ‡Poor Risk  §High Risk  ¶Very High Risk
expected probabilities of mortality for each of the five categories of patient risk proposed by Parsonnet et al. (1989).

The median of the AROCs listed in Table 3.5 is 0.74 with interquartile range of 0.72 to 0.76 and their mean is 0.75. The AROCs range from a minimum of 0.65 for the assessment of the Parsonnet score by Gabrielle et al. (1997), where the study data were collected from patients who underwent surgery at one of 42 cardiac surgical centres in France, to a maximum of 0.90 for the assessment by Vanagas et al. (2003), where the patients were treated at only one hospital. As discussed in Section 3.3.2, it is assumed that unmeasured hospital effects are one reason for the reduced predictive power of risk scores, which use patient characteristics as risk factors, when they are applied to populations of patients from more than one institution.

The average of the observed and expected probabilities of mortality are available for seventeen of the nineteen items listed in Table 3.5. The standardized mortality rate (SMR), defined as the ratio of the sum of the observed mortality outcomes and the sum of expected probabilities of mortality, is less than one for each study, less than 1/2 for nine of the studies, and less than 1/3 for four of the studies. Although the information from the studies in the literature review is insufficient to provide confidence intervals for the SMRs, there is an indication that the original Parsonnet score overpredicts the probability that a patient will die following cardiac surgery.

3.4.2 EuroSCORE

As noted in Section 3.3.2, the mortality response variable for the construction of the EuroSCORE was defined as death within 30 days from operation or, if the patient was undischarged 30 days after the operation, death in-hospital (Nashef et al., 1999). Only two of the twelve studies listed have this definition for mortality. For the others, six define mortality as in-hospital death, two as death within 30 days of the operation, and, for the study by Nashef et al. (2002) who used the STS database, the definition is operative mortality. Data for the EuroSCORE assessments were collected from patients treated at a single hospital in eight of
Table 3.6: Number of data sources (#Centres), areas under receiver operating characteristic curve (AROC), data set mean of expected probabilities of mortality (Expected) and observed mortality rates (Observed), and standardized mortality rates (SMR) from studies assessing the EuroSCORE.

<table>
<thead>
<tr>
<th>Study</th>
<th>Surgery</th>
<th>#Centres</th>
<th>AROC</th>
<th>Expected</th>
<th>Observed</th>
<th>SMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEI:00</td>
<td>Bypass Surgery</td>
<td>1</td>
<td>0.79</td>
<td>0.070</td>
<td>0.040</td>
<td>0.571</td>
</tr>
<tr>
<td>PIT:00/1</td>
<td>CABG &amp; Valve</td>
<td>1</td>
<td>0.81</td>
<td>0.032</td>
<td>0.020</td>
<td>0.625</td>
</tr>
<tr>
<td>PIT:00/2</td>
<td>CABG &amp; Valve</td>
<td>1</td>
<td>0.77</td>
<td>0.036</td>
<td>0.011</td>
<td>0.306</td>
</tr>
<tr>
<td>KAW:01</td>
<td>Bypass Surgery</td>
<td></td>
<td>0.82</td>
<td>0.053</td>
<td>0.045</td>
<td>0.849</td>
</tr>
<tr>
<td>SER:01</td>
<td>CABG</td>
<td>1</td>
<td>0.83</td>
<td>0.050</td>
<td>0.039</td>
<td>0.780</td>
</tr>
<tr>
<td>NAS:02/1</td>
<td>Cardiac Surgery</td>
<td>&gt;1</td>
<td>0.77</td>
<td>0.042</td>
<td>0.042</td>
<td>1.000</td>
</tr>
<tr>
<td>NAS:01/2</td>
<td>Cardiac Surgery</td>
<td>&gt;1</td>
<td>0.77</td>
<td>0.040</td>
<td>0.040</td>
<td>1.000</td>
</tr>
<tr>
<td>STO:02/1</td>
<td>Bypass surgery</td>
<td>1</td>
<td>0.86</td>
<td>0.052</td>
<td>0.043</td>
<td>0.827</td>
</tr>
<tr>
<td>STO:02/2</td>
<td>Bypass surgery</td>
<td>1</td>
<td>0.86</td>
<td>0.045</td>
<td>0.041</td>
<td>0.911</td>
</tr>
<tr>
<td>ASA:03/2</td>
<td>CABG</td>
<td>2</td>
<td>0.76</td>
<td>0.029</td>
<td>0.033</td>
<td>1.138</td>
</tr>
<tr>
<td>BRI:03</td>
<td>CABG</td>
<td>4</td>
<td>0.75</td>
<td>0.030</td>
<td>0.017</td>
<td>0.567</td>
</tr>
<tr>
<td>HUI:03</td>
<td>CABG &amp; Valve</td>
<td>1</td>
<td>0.84</td>
<td></td>
<td>0.023</td>
<td></td>
</tr>
<tr>
<td>KAR:03</td>
<td>Bypass surgery</td>
<td>1</td>
<td>0.82</td>
<td>0.037</td>
<td>0.012</td>
<td>0.324</td>
</tr>
<tr>
<td>VAN:03</td>
<td>Cardiac Surgery</td>
<td>1</td>
<td>0.79</td>
<td>0.078</td>
<td>0.056</td>
<td>0.718</td>
</tr>
<tr>
<td>NIL:06</td>
<td>Cardiac Surgery</td>
<td>1</td>
<td>0.82</td>
<td>0.051</td>
<td>0.029</td>
<td>0.569</td>
</tr>
</tbody>
</table>
the studies, from patients treated at more than one hospital in three studies, and Kawachi et al. (2001) do not give the number of hospitals participating in their study. Data were collected from patients undergoing isolated CABG surgery only in three studies and from patients undergoing a variety of cardiac-surgical procedures, including isolated CABG operations, in nine studies.

The AROCs of the EuroSCORE assessments in Table 3.6 range from 0.75 for the study by Bridgewater et al. (2003) to 0.86 for the two studies by Stoica et al. (2002). The median of AROC values is 0.81 with interquartile range from 0.77 to 0.825 and the mean is 0.80. It was noted in Section 3.4.1 that the median of the AROCs of the Parsonnet score studies is 0.74 with interquartile range from 0.72 to 0.76. That is, there is some evidence that the discriminatory power of the EuroSCORE is better than that of the Parsonnet score. Another important observation is that AROC of 0.79 found for the EuroSCORE when it was constructed is not significantly different from the average AROC of 0.80 for these EuroSCORE assessments.

The average predicted probabilities of mortality from the EuroSCORE are given for fourteen of the fifteen analyses in Table 3.6. The SMR for each study is less than one for 11 of the 14 studies and less than 1/3 for 2 of the studies. There is no study for which the SMR is in the interval (1/3, 1/2). It appears that, in general, the EuroSCORE overpredicts the probability that patients undergoing cardiac surgery will die, but the magnitude of the overprediction is less than that of the Parsonnet score.

### 3.5 Cardiac Surgical Risk Scores for a Single Institution

It appears from the analyses in Sections 3.3 and 3.4 that each of the models reviewed control adequately for the known patient characteristics that affect the risk of postoperative mortality. The variability in predicted probabilities of mortality seen in Section 3.4 after generalization the Parsonnet score or EuroSCORE must be due to one or more of the other sources of variation in the outcome identified by Lilford et al. (2004). The sources they identified were components of the case-mix
that cannot be measured, data definitions, data quality, clinical quality of care, and chance. There were no publications found that identified any institutional factors associated with variation of data quality or measured their effect on the performance of a cardiac surgical risk score at multiple cardiac surgical centres.

In a study on the quality of data used for monitoring processes in emergency medical systems, Dick et al. (1999) found that the data collected by different emergency services are often incompatible because each service differs significantly in structure and resources. They concluded the emergency service data should not be used in a quality improvement program as an information source for comparing performance of the emergency services in their treatment of trauma victims. It is probable, given the results of the review in Section 3.4, that there are similar issues with data collection by cardiac surgical teams and institutions. For example, there were three definitions of postoperative mortality—in-hospital, within 30 days of the operation, and in-hospital or within 30 days of the operation if the patient was discharged—in the publications comparing performance of the EuroSCORE and Parsonnet score. This issue of variation in data definitions and quality suggests that currently available cardiac surgical models, which will use patient characteristics only to risk adjust for the expected probability of mortality, should not be used for the purpose of comparing the quality of cardiac surgical care provided by hospitals or surgeons because the risk adjustment cannot guarantee that like is being compared with like (Lilford et al., 2004).

A surveillance program with a model that adjusts the expected probability of mortality using measures of the patients’ characteristics as risk factors is feasible, provided it is used to monitor the deaths at a single institution. The reason is that a risk model needs to be adapted to the measured mortality rate at the hospital of interest. In this section, two strategies to adapt cardiac surgical risk models to the hospital in which they are being used are suggested. In Section 3.5.1, a procedure to recalibrate an “off the shelf” risk score is discussed. An alternative, which is proposed in Section 3.5.2, is that an institution use its data to construct a customized cardiac surgical risk score to provide unbiased estimates to the probability of mortality following operations undertaken by its surgeons.
3.5.1 Recalibrated “Off the Shelf” Scores

The issue identified in Section 3.4 is that the calibration of the Parsonnet score or EuroSCORE is uncertain if either is used to stratify the risk of patients in populations remote from those used in its construction. Peterson et al. (2000) used the logistic model

$$\text{logit}(p^*) = \alpha + \beta(RS), \quad (3.1)$$

where $p^*$ is the revised predicted probability of mortality and RS is the original cardiac surgical risk score, to recalibrate cardiac risk models for the patients in their database. Steiner et al. (2000) used a training data set of 2,000 adult patients who underwent cardiac surgery at a single institution to estimate the parameters $(\alpha, \beta)$ of Equation (3.1). They then proposed the risk adjusted CUSUM scheme, in which the expected probabilities of mortality were calculated by substituting their parameter estimates and the patients’ Parsonnet scores into Equation (3.1), to prospectively monitor the postoperative mortality of patients undergoing cardiac surgery at that institution. In general, any currently available cardiac surgical risk model may be recalibrated to provide unbiased predictions of the risk of mortality for the patients undergoing surgery at a particular institution and the recalibrated risk score may be used in a scheme to monitor the outcomes of cardiac surgery undertaken at that institution.

3.5.2 Customized Scores

When constructing cardiac surgical risk scores, it appears many surgeons believe that it is possible to make fair comparisons of providers if an exhaustive array of surgical risk factors are included in a statistical model. For example, as noted in Section 3.3.2, the number of risk factors used in the STS model increased as the number of patients in the training data increased. Tu et al. (1997) studied the effect of the number of risk factors on the predictive power a risk model by adding covariates to the Ontario model (Tu et al., 1995). Its six risk factors are age categorized at three levels, female gender, emergency operation, previous CABG, measures of left ventricular function, and the presence of left main disease. It had an AROC of 0.77 for CABG patients which only increased to 0.79 after the addition of the risk.
factors recent myocardial infarct, Canadian Cardiovascular Society class 4 angina, peripheral vascular disease, cerebral vascular disease, and chronic obstructive pulmonary disease to the model. That is, the Ontario model has risk factors from the Age, Unclassified, Diagnosis, Cardiac Failure, and Reduced Blood Supply to the Heart classes of meta-factors. The addition of more risk factors from the Reduced Blood Supply to the Heart class and some risk factors from the Comorbidity class of meta-factors resulted in a marginal increase in the predictive power of the model. The investigation by Tu et al. (1997) also showed that most of the risk of postoperative patient mortality was explained by age, emergency surgery, previous CABG, and left ventricular function, where age explained the most risk and left ventricular function the least. They concluded that simpler models may be just as effective as more complex models for interprovider comparisons of the short-term mortality risks from CABG.

The review of cardiac surgical models for the risk of postoperative mortality in Sections 3.2 and 3.3 supports a conclusion that simpler models may be as effective as more complex models for risk adjustment of cardiac surgical patients. All models reviewed use risk factors to stratify the patients’ risk of mortality. Classification of the risk factors into the meta-factors proposed in Section 3.3.1 clearly defines the risk factors in each model as measures of physiological reserve, cardiac reserve, or diagnosis. In addition, almost half the models use risk factors which are a measure of acute illness, to predict the patients’ risks of adverse postoperative outcomes. The number of risk factors the models use ranges from six (Tu et al., 1995; Staat et al., 1999) to forty-six (Gabrielle et al., 1997), but the scatter plot of their AROCs and number of risk factors in Figure 3.1 indicates that a higher number of covariates in cardiac surgical risk models does not necessarily mean an increased ability to stratify patients’ postoperative risk of mortality.

There is evidence in the literature supporting the conclusion that models with a limited number of risk factors are effective for risk adjusting cardiac surgical patients. Therefore, it is feasible that an institution construct a customized risk score using data from its patients. Such a strategy has been used for risk adjusting ICU patients. For example, Graham and Cook (2004) used a training and a test data set with 3,708 and 1,570 patients, respectively, to construct a risk model.
which used the following five covariates to predict the probabilities of mortality of intensive care patients admitted to the Princess Alexandra Hospital in Brisbane, Australia:

- patient’s age in years;

- lead time, which is a categorical variable coded < 1d for ICU admission the same day of hospital admission, 1d for ICU admission the day after hospital admission, 2d for admissions up to 2 days after hospital admission, and > 2d for all other ICU admissions;

- disease risk category coded low, neutral, or high, where the coding for both non-operative and operative, major disease groups is given by Graham and Cook (2004, Tables 2 and 3);

- comorbidity score given by the Chronic Health Evaluation component of the APACHE III score (Knaus et al., 1991); and

- acute physiology score given by the Acute Physiology component of the APACHE III score.

They show that the performance of their model is comparable with that of the APACHE III score, which is a more complex model for the risk of mortality of patients admitted to intensive care units. It has been shown to be poorly calibrated if used to predict the expected probabilities of mortality of patients admitted to ICUs in centres that did not contribute data to its developmental database (Cook, 2000).

In another study, Mullany et al. (2004) constructed a model which used three risk factors to predict the risk of death of patients admitted to the paediatric intensive care unit (PICU) at The Prince Charles Hospital in Brisbane, Australia. The risk factors used were (1) diagnostic category with low, intermediate, and high risk assigned according to the procedure performed, (2) weight coded \( \leq 2.5 \)Kg, > 2.5 and \( \leq 3.5 \)Kg, and > 3.5Kg, and (3) log(age) where the infant’s age is in days. Their model was developed using a data set of 2,156 admissions to the PICU. They found its AROC value of 0.83 was not statistically different from the AROC
value of 0.90 for the Paediatric Index of Mortality 2 (PIM2) (Slater et al., 2003), which is the “gold standard” for benchmarking and performance assessment (of PICUs). PIM2 uses the nine risk factors and the formula given by Slater et al. (2003, Appendices) to calculate the expected probability of mortality.

Clearly, this process to construct customized ICU scores could be adapted for cardiac surgical scores by cardiac surgical teams who wish to monitor patient outcomes as part of a quality improvement program. Such scores could save the time and expense of data collection by using the appropriate, routinely-collected data for the risk factors. For adequate predictions of patients’ risk of postoperative mortality, it is necessary that the risk factors in these customized models provide at least one measure for each of the dimensions of risk for mortality following cardiac surgery discussed in Section 3.3.1. Given the finding by Tu et al. (1997) that the predictive power of their cardiac surgical models with more than six risk factors was marginally greater than that of their model with only six risk factors, it is assumed that the performance of customized models would not be greatly improved if there were more than six risk factors. Using the well known rule of thumb, discussed in detail in Chapter 6.5.1, a model with six risk factors would require 70 deaths (10 per estimated parameter) in its training data set. That is, if the observed mortality rate of the cardiac surgical population was one per cent then the expected size of the training data set would be 7,000.

Provided the model fit was undertaken with data collected when the surgical process was predictable, the customized model may be expected to provide unbiased estimates of the probability of mortality of adult cardiac surgical patients until the institution’s processes are changed. If any process improvements are implemented, it will be necessary to refit the model for unbiased estimates of the probability of postoperative mortality for in-control processes.

### 3.6 Conclusion

This review of adult cardiac surgical risk scores was in two parts. In the first, the risk factors and predictive power of 23 models were compared. There were over a hundred measures of patient characteristics used as risk factors, but they could
be classified into eight meta-factor categories. Such a classification showed that each model had at least one risk factor as a measure of the patients’ physiological reserve and diagnosis and all but one had at least one risk factor as a measure of the patients’ cardiac reserve. Although the risk factors used as measures of patient characteristics may vary according to the risk model, it appears that the underlying patient characteristics used to adjust the patients’ level of risk remain relatively constant. It is probable that this is the reason that the predictive powers of all models were comparable.

Furthermore, the analysis in Section 3.3 indicates that the predictive power of each cardiac surgical risk model appears to be independent of the number of risk factors used in its construction. A possible reason is that each of the risk factors in a model is a measure for one of a limited number of patient characteristics which affect the risk of mortality of cardiac surgery. Accordingly, a hypothesis, that one or two risk factors for each patient characteristic provide a sufficient measure of that characteristic, was proposed in Section 3.5.2. The study by Tu et al. (1997), which shows that the addition of more than two risk factors for a particular patient characteristic to their model only gave a slight improvement of the model’s predictive power, with the AROC increasing from 0.77 to 0.79, supports this hypothesis.

The concept of meta-factors has proved useful for understanding the performance of cardiac surgical risk models. However, further work, such as consultation with expert cardiac surgeons or cardiologists, is needed to confirm that the meta-factors and dimensions of risk, as proposed, are truly adequate for quantifying the risk of mortality following cardiac surgery.

Each of the risk scores reviewed in this chapter has been developed from a single database of patients. There has been debate as to whether they can be generalized to other practices or other countries (Gogbashian et al., 2004). The summaries of studies assessing the performance of the Parsonnet score and the additive EuroSCORE in Section 3.4 show that, if these scores are generalized to patients in databases not used for their development, both scores retain their ability to stratify the level of patient risk, but both tend to overestimate the probability of patient mortality. The summaries in Tables 3.6 and 3.5 indicate
that, although both the EuroSCORE and the Parsonnet score tend to overestimate the risk of postoperative mortality, the EuroSCORE tends to be more robust than the Parsonnet score in that it generally has better discrimination of the patients’ risk of mortality and its overestimation of the probability of mortality is less than that of the Parsonnet score.

The results from this literature review support the findings of Gogbashian et al. (2004) who reviewed individual centre and regional studies to assess the international performance of the additive EuroSCORE. They concluded that it overestimates the probability of mortality for scores $\leq 6$ and underestimates for scores $> 13$. If accurate predictions of the probability of mortality are required, neither the Parsonnet score nor the EuroSCORE should be generalized to patient populations remote from their developmental databases without careful calibration to ensure that the predicted probabilities of mortality provided by the risk scores match the observed mortality rates.

The chapter concluded with a proposal that data from patients treated at an institution may be used to recalibrate the risk score so that the risk adjustment for patients undergoing surgery at that institution is unbiased. Alternatively, a hospital may use its data to construct a customized risk model using a limited number of risk factors appropriately chosen from the meta-factor classes. It is clear that such a customized score will provide an institution with estimates of the patients’ risk that are at least comparable to the estimates from any “off the shelf” score.
Chapter 4

Multi-Institutional Monitoring Schemes for Adverse Hospital Events

4.1 Introduction


Some measures of hospital performance included in the Queensland Health report card are:

- number of patients who underwent elective surgery during the year;
- number of patients waiting to undergo elective surgery at the end of the year;
- indices of patient satisfaction;
• proportion of staff on sick leave;
• measures of clinical performance; and
• patient cost per weighted separation.

Queensland Health defined a separation as a process by which an admitted patient completes a process of care. Cost weights are used to adjust a hospital’s activity to compensate for the number and complexity of cases treated. The use of weighted separations allows for more accurate comparison of hospitals with specialist services with general hospitals where less complex cases are treated.

Clinical performance was assessed by comparing a hospital’s adverse outcomes rates following surgery or treatment for medical conditions such as stroke or acute myocardial infarction (AMI) with the average outcome rate of all hospitals in its so-called “peer group”. Peer grouping categorized the 42 Queensland public hospitals in the report card into three groups of hospitals; namely, (i) principal referral and specialized, (ii) large, and (iii) medium and small, where the numbers of hospitals in the respective peer groups are 10, 14, and 18. The hospitals in each peer group were of similar size and provided similar types and volumes of services (Queensland Health, 2007). It is stated that the quality measures of peer grouped hospitals better represent the differences in patient views and hospital performance and not just the differences in the range of services that the hospitals provide. Peer grouping is said to allow valid comparisons of within peer group hospital performance.

Each hospital was assessed by comparing its risk adjusted outcome rates and associated 99.9% confidence intervals with the state and peer group medians. Each risk adjusted outcome rate is derived by first calculating the expected outcome rate for each hospital, where the expected outcomes for the hospital’s patients were computed from a fitted logistic model with age, sex, and some comorbidities as risk factors. The criteria used for selecting a comorbidity as a risk factor for an outcome of interest included its frequency of occurrence in the treatment group, specialist medical advice, its statistical relationship with the outcome, and any available evidence from the literature. Next, the risk adjusted outcome rate was found as the ratio of the observed and expected outcome rates multiplied by the
total rate of outcomes per 100 separations for the entire cohort. The clinical report card shows that 23 of over 500 outcome measures showed significant variation (at the 99.9% confidence level) from the peer group and/or state average.

An issue with using report cards to compare hospitals is that, in spite of caveats that “differences in performance may occur as a result of a number of factors”, any measure of adverse patient outcomes that is higher than the average is taken as an indication of poor performance. For example, on 7 March 2007, an article in the Courier Mail newspaper (Miles, 2007) stated that the heart attack victims admitted to the Redcliffe hospital had a mortality rate almost double the state average, although the report card showed that the hospital’s mortality rate was not significantly different from the state or its peer group average. Furthermore, the newspaper reported that the Queensland president of the Australian Medical Association stated the report was proof that much more work was needed in the public hospital system, although neither the proof nor the type of work to which she was referring was reported.

This chapter assesses publication of outcomes in report card form. It commences with a review in Section 4.2 of the literature on the debate about the use of report cards as a quality improvement tool. The New York State Department of Health pioneered the use of report cards to publish hospital mortality rates following cardiac surgery. Section 4.3 contains a detailed description of the New York State report card, a literature review of previous assessments of it, an analysis of the proportion of “outlier” hospitals it reports, and a proposal for an alternative method for publishing mortality rates following cardiac surgery in New York State. In Section 4.4, data for the predicted probabilities of mortality and outcomes for acute myocardial infarct patients admitted to eleven hospitals in Yorkshire, England, are analysed using box plots, with fixed and random effects models, and risk adjusted CUSUM schemes to identify “outlier” hospitals. The results show that the hospitals which are identified depend to some degree on the analytical method used. The chapter concludes in Section 4.5 with some discussion of the need for clinicians to be involved in report card development.
4.2 Literature Review of the Debate about the Use of Report Cards

Publication of report cards to improve the performance of health care institutions and medical practitioners has been justified using several arguments. For example, report cards lead to informed choices by enhancing the consumers’ ability to compare health care providers (Hibbard et al., 2000; Mukamel et al., 2004/2005). Disclosure of information on a surgeon’s performance allows the patients to make effective informed consent when agreeing to undertake a procedure (Clarke and Oakley, 2004). Report cards on outcomes allow managers to focus on the delivery of products and services (Baker and Porter, 2004). Clinical outcome measures that are presented in an easily understood format may be used as decision tools that aid health care purchasers to make rational decisions when purchasing health care plans (Hibbard, 1998). It is argued that publication of performance indicators will lead to performance improvement (Mullen, 2004).

Academics and medical practitioners have criticized the report cards currently published for many perceived shortcomings. There have been issues raised about the quality and timeliness of the data used for the report cards (Goldstein and Spiegelhalter, 1996; Leyland and Boddy, 1998; Simon, 2001; Mitchell et al., 2004; Rowan et al., 2004). Medical practitioners believe that risk adjustment to remove the variation due to patient heterogeneity is necessary before any comparison of medical practitioners or institutions is valid (Iezzoni, 1997).

However, past and current risk adjustment models are seen to be inadequate because they do not control for all of the differences between patients or institutions. For example, a hospital which treats patients from a typical sample drawn from residents of a nursing home would have a higher than normal share of patients over 85 years should not be compared with an institution which treats patients drawn from the general population, unless the risk adjustment for age is adequate (Green et al., 1991). Risk models do not allow for variation due to the location of an institution (Seagroatt and Goldacre, 2004) and, because they have a limited number of explanatory variables, there is the possibility that some of the variation
is due to unknown confounding factors (Mitchell et al., 2004). In general, the risk adjustment is seen to be imperfect (Goldstein and Spiegelhalter, 1996; Leyland and Boddy, 1998; Poses et al., 2000; Rabilloud et al., 2001; Sharif and Afnan, 2003; Editor, 2004; Rowan et al., 2004) because the model does not remove all the uncertainty due to the patient mix and so the residual variation is due to factors other than the quality of care and random chance. If institutions being compared are assumed to be broadly similar, the appropriate statistical analyses using random effects models of the data used for the report cards (Goldstein and Spiegelhalter, 1996; Marshall and Spiegelhalter, 1998; Rabilloud et al., 2001; Denton et al., 2002) indicate, in most instances, that an institution’s ranking is due to chance. Thus, an investigator cannot infer from these reports that the quality of care and service provided by one institution is different from that provided by another.

Medical practitioners are concerned that report cards are used, by the press in particular, in a judgmental way (Marshall et al., 2002). The newspaper article by Miles (2007) in Section 4.1 provides a good example of the press passing judgment on the quality of care in public hospitals based on evidence provided by one report card only. Another of their concerns is the possibility of “gaming” the system. Gaming means that procedures are undertaken or data are collected in a way that will manipulate performance indicators showing the institution or medical practitioner in the best possible light, but will not be indicative of the quality of care provided (Lilford et al., 2004). They define gaming as a technique commonly in use to find loop-holes to reduce published mortality rates by improved data collection, or by intentional deception. Some examples of gaming strategies are given as:

- referring high risk patients to other medical centres (Omoigui et al., 1996);
- submitting gamed data if the request for information is perceived as ridiculous (Marshall et al., 2002);
- altering records (Pitches et al., 2003);
- inappropriately suspending patients from waiting lists (Pitches et al., 2003);
• removing a patient from a waiting list after an offer of an appointment at a time the patient is known to be going on holiday is declined (Pitches et al., 2003); and

• manipulating reporting criteria to improve the results reported by an in vitro fertilization clinic (Sharif and Afnan, 2003).

The stated intention of publishing report cards is that the public will use report cards to make more informed decisions about their medical treatment, but there are studies (Schauffler and Mirdavshy, 2001; Werner and Asch, 2005) indicating the public do not use them for that purpose. Surveys do reveal that public accountability for the standard of health care provided is important (Werner and Asch, 2005), so public health authorities have an incentive to publish performance reports. The New York State Health Department publishes risk adjusted outcomes for cardiac surgery (NY State, 1994-2004) with the stated purpose of achieving co-operative quality improvement. The National Health Service, Scotland, publishes outcomes with a focus of making data more influential in leveraging quality improvement (Clinical Outcomes Group, 2004). In 2002, eighteen organizations publicly reported 333 measures of health care quality in California (Broder et al., 2005).

Despite the concern that more research is needed to develop accurate and useful quality related performance measures (Marshall et al., 2003) and the other issues raised in the preceding paragraphs, public reports of quality performance measures will continue. There is, therefore, a need for clinicians to become involved in and to take control of the process of publishing performance measures (Tu et al., 2001). Ideally the publication of performance measures should be part of a continuous quality improvement (CQI) program implemented by the health care workers (Shahian et al., 2001) and that program should have ‘on line’ monitoring of outcomes and performance measures to detect in a timely way any fluctuations that may be attributed to special causes.
4.3 Report Cards for Cardiac Surgery in the New York State Health Department

The first instance of publication of the names of hospitals and their patients’ death rates occurred in the United States in 1986 (Brinkley, 1986). At first, the Health Care Financing Administration (HCFA) was compelled to release that information to the New York Times newspaper under the Federal Freedom of Information Act, but, subsequently, the HCFA and others in possession of such data voluntarily released hospital mortality rates (Green, 1992). For example, when releasing details of patient mortality after admission to hospital, the HCFA held press conferences and distributed press kits which highlighted any high mortality “outlier” hospitals.

The New York State Department of Health responded to the shift in the focus of the quality of care in US hospitals to the analysis of patients’ outcomes by establishing a Cardiac Advisory Committee composed of practising New York cardiac surgeons and cardiologists and advisors from other US states. Its purpose was to investigate issues related to quality assurance, appropriateness of surgery, and prevention and control of cardiovascular disease for cardiac patients in the state (Hannan et al., 1990). From 1988, hospitals certified by the Department of Health to undertake open heart surgery in New York State provided information on all of their cardiac surgical patients in the form of the Cardiac Surgery Reporting System (CSRS). The CSRS became the first profiling system with sufficient clinical detail to generate credible comparisons of providers’ outcomes and has been recognized by many states and purchasers of care as the gold standard among systems of its kind (Green and Wintfeld, 1995). This assessment, that the CSRS is a leading example among databases used to compare health care providers, merits careful evaluation (Green, 1992).

For this evaluation, the report card derived from the CSRS data is described in Section 4.3.1 and some of the issues raised in previously published assessments are discussed in Section 4.3.2. Reports on each hospital’s mortality rate for patients undergoing CABGs are available for the years 1994 to 2004 (NY State, 1994–2004). Some simple analysis of the mortality rates given in Section 4.3.3 indicates that
the proportion of hospitals with mortality rates that are significantly different from
the state average is twice the expected proportion for the stated 95% confidence
level. An alternative to using report cards for publication of the latest available
hospital or surgeon-specific mortality rates is to publish an annual trend analysis
of mortality rates in the form of time series charts for each hospital. Plots of the
annual mortality rates following CABG surgery for two New York State hospitals
are presented and discussed in Section 4.3.4.

4.3.1 Description of Monitoring Scheme Used by NY State Health

Report cards on adult cardiac surgery in New York State are published in the form
of booklets which contain some introductory remarks, the risk model used to pre-
dict probabilities of patient mortality following CABG surgery, and a number of
report cards giving risk adjusted mortality rates at the hospital or surgeon specific
level. It is stated in the introductory section of each report card that its objectives
are to provide information on risk factors associated with coronary artery bypass
and heart valve surgery and to list hospital and surgeon-specific mortality rates
which have been risk-adjusted to account for differences in patient severity of ill-
ness (NY State, 2006). The introduction continues with brief descriptions of and
possible treatments for coronary artery and heart valve disease and gives outlines
of the Health Department’s program, the patient population, and the risk adjust-
ment process. In this study, the focus is on the report cards for the mortalities
following isolated CABG surgery\(^1\) in each of the hospitals certified by the Depart-
ment of Health and the risk model used to calculate the expected probability of
mortality for patients undergoing CABG surgery.

*Hospital Report Card for Mortality Following CABG*

A report card for patient mortality following CABG surgery provides (i) the num-
ber of CABG cases who were discharged during a calendar year, (ii) the number of

\(^1\)Isolated CABG surgery refers to cardiac operations for coronary artery bypass grafts without
concurrent valve repair/replacement or any other cardiac surgical procedure.
SECTION 3: CSRS IN NEW YORK STATE

deaths, (iii) the observed mortality rate (OMR), (iv) the expected mortality rate (EMR), (v) the risk adjusted mortality rate (RAMR), and (vi) the 95% Confidence Interval (CI) for the RAMR of each hospital registered to undertake CABG surgery in New York State. These variables are defined by the New York State Health Department (NY State, 2006) as follows:

**Death** Prior to the 2004 report, mortality data were collected for in-hospital deaths, which were defined as patients who died subsequent to CABG surgery during the same admission, or died following discharge to hospice care. In the 2004 report, mortality data were collected for patients who died during the same hospital stay in which the patient underwent CABG surgery or discharged patients, including those patients discharged to a hospice, who died within 30 days of surgery.

**OMR** is the ratio of the number of deaths and the number of patients.

**EMR** is found by summing each patient’s expected probability of death predicted by the risk score and dividing that sum by the number of patients treated.

**RAMR** is a scaled form of the standardized mortality rate given by the formula

\[
\text{RAMR} = \frac{\text{OMR}}{\text{EMR}} \times \bar{p},
\]

where \( \bar{p} \) is the OMR for all patients who underwent CABG surgery in hospitals registered with the New York State Department of Health. It gives the best estimate of the provider’s mortality rate if the provider had a mix of patients identical to the statewide mix (NY State, 2006).

**95% CI for RAMR.** The method used to compute the CI is not given.

Any RAMR significantly above the state average is flagged with a single asterisk (*) and any significantly below the state average is highlighted by a double asterisk (**).

*Risk Adjustment Model*

The expected probability of mortality of each patient who underwent CABG
surgery is given by the multivariable risk factor equation for CABG Hospital Deaths in New York State (NY State, 1994–2004). It is constructed using backward stepwise regression to fit a logistic regression model to CSRS data collected from New York State hospitals (Hannan et al., 1990, 1994). The model is reconstructed each year by refitting the model to the CSRS data collected for that year. Thus, the explanatory variables and associated parameter estimates of the risk model vary from year to year, but the statewide EMR equals the statewide OMR in each of the annual report cards.

4.3.2 Issues with the New York State Health Report Cards

Hannan et al. (1994) state that the New York State Department of Health was forced to release surgeon-specific information for 1989 and 1990 to *Newsday* in 1991 after losing a lawsuit filed by *Newsday* on the basis of Freedom of Information Law. In December 1992, the Department voluntarily released surgeon risk-adjusted CABG mortality rates for the three year period 1989 through 1991. The publication of a report card on surgeons and its annual updates was an important and controversial new trend (Green and Wintfeld, 1995). Although the probability of mortality of patients who died following CABG surgery was risk adjusted, these reports indicated that the percentage of patients who died differed widely among surgeons, but risk-adjusted mortality rates for CABG surgery had declined in New York State following the implementation of CSRS. Thus, some concluded that such systems may save lives.

According to Green and Wintfeld (1995), these important findings were controversial because of scepticism, particularly among clinicians, about the validity of the CSRS data and methods. To help inform the debate about the proper role of report cards in the health care system, they and other authors conducted an evaluation of New York’s CSRS. The following summarizes some of the issues raised in that debate.

**Timely Publication**

Green and Wintfeld (1995) note that the degree of fluctuation of the surgeons’ per-
formance ratings from year to year means that, by the time the data are published, users of the report can have little confidence that the ratings are still applicable. Timely publication of the cardiac surgical reports appears to be an ongoing issue. The delays for the annual reports on cardiac surgery undertaken in the ten years 1995 to 2004 ranged from 18 to 28 months.

Quality of Data and Risk Factor Gaming

Green and Wintfeld (1995, Table 1) also identified an increase in the prevalence of five risk factors used in the CSRS models in each of the years 1989, 1990, and 1991. The risk factors identified were renal failure, congestive heart failure, chronic obstructive pulmonary disease, unstable angina, and low ejection fraction (< 40%). Green and Wintfeld note that the reported increases in 1990 were associated with the participating hospitals receiving the first report comparing their death rates in February 1990. The increases in 1991 were associated with the publication of a list of hospitals ranked according to their CABG-related mortality rate in December 1990.

It was hypothesized that the reason for these increases over time was that the surgeons who were responsible for the data collection became increasingly aware that they were the principal subjects of the investigation. Green and Wintfeld (1995) found that one hospital reported an increase from 1.8% to 52.9% for the prevalence of chronic obstructive pulmonary disease and another reported that unstable angina had increased from 1.9% to 60.6%. At the individual level, surgeons reported greater variations in the prevalence of these risk factors than could be reasonably attributed to differences in patient populations. The prevalence of chronic obstructive pulmonary disease reportedly varied from 1.4% to 60.6% and that of unstable angina ranged from 0.7% to 61.4%. On the other hand, Green and Wintfeld (1995) were unable to resolve questions about the accuracy and internal consistency of the CSRS database because the CSRS data collection instrument was substantially changed in 1991 (Hannan et al., 1994).

These apparently spurious increases in risk factors accounted for 66% of the increase in the state-wide predicted probability of mortality which was 2.62% in 1989 and 3.16% in 1991 (Hannan et al., 1994, Table 1). Green and Wintfeld
(1995) concluded that these unexplained increases in the expected probability of mortality were responsible for 41% of the reported reduction in the state-wide RAMR from 4.2% in 1989 to 2.7% in 1991. Because a proportion of the reduced RAMR is attributed to improved data collection, it is given as an example of gaming (Burack et al., 1999; Lilford et al., 2004). In a survey where 104 of 150 New York State cardiac surgeons responded, 40% of the respondents cited risk factor gaming as an aspect of CSRS needing most improvement (Burack et al., 1999, Table 2). They give as examples of gaming in the operating room the insertion of a prophylactic intra-aortic balloon pump before induction and making a CABG into a CABG/ventricular aneurysmorrhaphy with one or two additional sutures. An example of gaming in the postoperative phase was the practice of transferring patients, who have sustained massive perioperative neurologic injury, to a chronic care facility. According to Burack et al. (1999) such transfers are difficult to track, so the “decanting” of moribund patients is commonplace and can favourably affect hospital mortality rates.

There is, however, a process to manage data entry into the CSRS. Hannan et al. (1994) reported that data quality checks were contained in the database package and the data for 1989 and 1992 were subjected to a quality audit. In the 1992 audit, CSRS data were extracted from 50 randomly chosen, medical records of CABG surgery patients from each of ten hospitals. The audit team found that the average expected mortality of all their cases was not significantly different from that calculated from the coding entered by the hospitals. They concluded the risk factors entered into CSRS by the hospital staff were of sufficient accuracy. The audit process is ongoing. It is stated in the latest cardiac surgery report (NY State, 2006) that data are verified through a review of unusual reporting frequencies, cross-matching of cardiac surgery data with other Department of Health databases, and a review of medical records for a selected sample of cases.

**Predictive Accuracy**

Green and Wintfeld (1995) concluded that the predictive capacity of the CSRS model and the usefulness of the risk adjusted data were limited. They found that changes in surgeons’ rankings during the two years of the study were so
extensive that in one year 46% of the surgeons had moved from one half of the ranked list to the other. Burack et al. (1999) were also critical of the CABG predictive mortality models in general and the CSRS models in particular. They questioned the predictive power of the CSRS models by comparing the measure of their predictive power, the $C$-index (Agresti, 2002, Chapter 6.2.6), with the $C$-index of weather models. They stated the respective $C$-indices ranged between 0.79 and 0.81 and between 0.71 and 0.89.

Hannan et al. (1990) compared the risk model using the CSRS data with models using discharge or claims data and found that the CSRS model provided significant improvements in predictive ability. A goodness of fit test (Hannan et al., 1994) confirmed the CSRS model predicts mortality with reasonable accuracy and there was a reasonably good correspondence between actual deaths and expected deaths for all levels of risk. There is no indication that the Department of Health or the Cardiac Advisory Committee intended to use the surgeons’ rankings in the report cards to judge their performance. The goal, as stated in the report cards, is to provide hospitals and cardiac surgeons with data about their own outcomes for cardiac procedures which allows them to examine the quality of care they provide and to identify areas that need improvement.

Refusal of Treatment
Two New York newspapers provided anecdotal evidence that cardiac surgeons were refusing to treat high risk patients (Zinman, 1991; Byer, 1993). The source for the report by Zinman was “many doctors”. It was reported severely ill heart patients were finding it increasingly difficult to get surgery because some surgeons and hospitals were refusing to take patients because they feared that the death of high risk patients would lower their standing in state mortality statistics.

Burack et al. (1999) reported that 67% of the 104 respondents to their survey had refused treatment to at least one high risk CABG patient in the last year and that patients with an ascending aortic dissection, whose outcomes are not reported, were more likely than CABG patients to be accepted for treatment. However, an analysis of data for Medicare patients who underwent CABG surgery (Peterson et al., 1998) did not provide any empirical evidence that access to care for New
York bypass patients had declined. In a recent study of cardiac surgical practices in northwest England, Bridgewater et al. (2007) also concluded that there was no evidence that fewer high risk patients are undergoing surgery because mortality rates are published.

**Migration out of state**

After a retrospective analysis of isolated CABG operations undertaken at the Cleveland Clinic (in the state of Ohio), Omoigui et al. (1996) concluded that public dissemination of outcome data may have been associated with increased referral of high-risk patients from New York State to an out-of-state regional medical centre. In contrast, an analysis of a national database of Medicare patients aged more than 64 years (Peterson et al., 1998) provided no evidence for a widespread increase in out-of-state transfers for by-pass surgery.

**Questionable Decline in Mortality**

Hannan et al. (1994) reported a decrease in the actual mortality rate of CABG patients undergoing surgery in New York State in the years 1989 to 1992 despite the average severity of illness of the patients increasing over that period. As a result, the RAMR decreased from 4.1% in 1989 to 2.45% in 1992. However, Green and Wintfeld (1995) have questioned the reductions in mortality because it was possible that they were the result of (a) refusing treatment for high risk patients, (b) transferring them out-of-state, or (c) gaming risk factors. Finally, they note that reductions in mortality in New York State must be interpreted in the light of the evidence of a nationwide reduction in mortality rates in elderly patients who had angioplasty or bypass surgery from 1987 to 1990 (Peterson et al., 1994).

Bridgewater et al. (2007) also found that the mean observed mortality rate and mean expected probability of mortality of cardiac surgical patients undergoing operations in north-west England between 1997 and 2003 decreased and increased, respectively, over time. In another recent study using data extracted from all admissions for CABG in England between 1996-7 and 2003-4, Aylin et al. (2007) found that the year of the operation was associated with a reduction in the risk of postoperative mortality. A possible interpretation of such findings is that, as
a professional group, cardiac surgeons are improving their surgical technique and, consequently, their patients’ chances of surviving the operations as the number of patients who have undergone cardiac surgery increases. Such learning curves were first described in the aircraft construction industry where it was found that the number of errors made in constructing an aircraft decreased as the production numbers of that aeroplane increased (Wright, 1936).

**Questionable Effectiveness in Informing Patients’ Decisions**

In 2004, Bill Clinton, the ex-President of the United States, underwent CABG surgery at the Columbia Presbyterian hospital which, according to the 2001 report card that was the latest available at the time of the report, had the highest mortality rate for that operation in New York State. The New York Times (Altman, 2004) reported that Mr Clinton was awaiting surgery and raised the issue of the effectiveness of New York State report cards to help better inform patients in deciding where they want to go for CABG operations. The article stated that, like many others who suddenly learn they need bypass surgery, Mr Clinton had little chance to study statistics or consult experts. He simply found himself on a medical-referral track that led to Columbia Presbyterian.

The sentiments expressed in the New York Times article have been echoed in other publications (McMahon, 2004; Baxter, 2005). However, as noted in Section 4.2, subsequent research indicates that, even if public report cards have a limited impact on improving health care quality, they fill another important need. It is that publication of report cards on the quality of health care allows the public to hold the health care providers accountable for their quality of health care (Werner and Asch, 2005). There is further discussion of this issue in Section 4.5.

### 4.3.3 Overdispersion of the Risk Adjusted Mortality Rates

In their first study using CSRS data, Hannan et al. (1990) identified seven of 28 hospitals with observed mortality rates which were different from the expected mortality rates at the 95% confidence level. There were three hospitals with a lower than expected observed mortality rate and four with a higher than expected
mortality rate, but Hannan et al. (1990) noted that only one high and one low outlier would be expected based on chance alone. In another study using the CSRS database, Hannan et al. (1994, Table 4) showed that ten of the thirty hospitals in the study were outliers in 1989. That study also showed nine, four and two outlier hospitals in New York State in 1990, 1991, and 1992, respectively.

Table 4.1 gives the number and proportion of hospitals which were flagged in each of the years 1994 to 2004 as having RAMRs significantly different, at the 95% confidence level, from the state average. This table shows that, for each of the nine years from 1994 to 2002, 9.7% or more hospitals were reported as “outliers”. If the variation in the RAMRs were due to chance alone, one would expect that approximately 5% of the New York State hospitals would be reported as outliers in any year. That is, the New York State monitoring system consistently reported more than the expected number of hospitals with RAMRs significantly higher or lower than the state average. It is possible, therefore, that the number of false alarms is excessive and, consequently, the number of investigations of hospitals with cardiac surgical processes flagged as out of statistical control is also excessive.

In 2003 and 2004 the proportion of reported outlier hospitals is less than 5%. One possible reason for the sharp drop in the number of reported outlier hospitals is that there is less variation in the hospitals’ interpretation of the definition of mortality following cardiac surgery because its definition for the CSRS database was changed after 2002 (see Section 4.3.1). Another possible explanation is the negative publicity the Columbia Presbyterian hospital received when Bill Clinton underwent heart surgery (see Section 4.3.2). The 2003 and 2004 report cards were published in 2005 and 2006, respectively, well after the negative report in the New York Times in 2004. It is possible that the New York State hospital administrations have responded to the report by ensuring the collection of their cardiac-surgical data adheres to the CSRS guidelines.

The data and information used for this assessment of the report cards for Adult Cardiac Surgery in New York State are freely available on the internet and, as stated in Section 4.3.1, the method for computing the confidence intervals was not given. Clearly, the method used for determining the 95% confidence intervals is an important question, but it will require liaison with the New York State Health
Table 4.1: Number and percentage of New York State hospitals with mortality rates following CABG surgery significantly below, above, and the total significantly different from the state average for the eleven years 1994–2004. The confidence level is 95%.

<table>
<thead>
<tr>
<th>Year</th>
<th>N*</th>
<th>Above n\textsuperscript{†}</th>
<th>Above %</th>
<th>Below n\textsuperscript{†}</th>
<th>Below %</th>
<th>Total n\textsuperscript{†}</th>
<th>Total %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>31</td>
<td>2</td>
<td>6.4</td>
<td>1</td>
<td>3.2</td>
<td>3</td>
<td>9.7</td>
</tr>
<tr>
<td>1995</td>
<td>31</td>
<td>1</td>
<td>3.2</td>
<td>2</td>
<td>6.4</td>
<td>3</td>
<td>9.7</td>
</tr>
<tr>
<td>1996</td>
<td>32</td>
<td>1</td>
<td>3.1</td>
<td>3</td>
<td>9.4</td>
<td>4</td>
<td>12.5</td>
</tr>
<tr>
<td>1997</td>
<td>33</td>
<td>3</td>
<td>9.1</td>
<td>1</td>
<td>3.0</td>
<td>4</td>
<td>12.1</td>
</tr>
<tr>
<td>1998</td>
<td>33</td>
<td>2</td>
<td>6.1</td>
<td>2</td>
<td>6.1</td>
<td>4</td>
<td>12.1</td>
</tr>
<tr>
<td>1999</td>
<td>33</td>
<td>2</td>
<td>6.1</td>
<td>1</td>
<td>3.0</td>
<td>3</td>
<td>9.1</td>
</tr>
<tr>
<td>2000</td>
<td>34</td>
<td>2</td>
<td>5.9</td>
<td>3</td>
<td>8.8</td>
<td>5</td>
<td>14.7</td>
</tr>
<tr>
<td>2001</td>
<td>35</td>
<td>2</td>
<td>5.7</td>
<td>2</td>
<td>5.7</td>
<td>4</td>
<td>11.4</td>
</tr>
<tr>
<td>2002</td>
<td>36</td>
<td>3</td>
<td>8.3</td>
<td>3</td>
<td>8.3</td>
<td>6</td>
<td>16.7</td>
</tr>
<tr>
<td>2003</td>
<td>37</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>2.7</td>
<td>1</td>
<td>2.7</td>
</tr>
<tr>
<td>2004</td>
<td>39</td>
<td>1</td>
<td>2.6</td>
<td>0</td>
<td>0.0</td>
<td>1</td>
<td>2.6</td>
</tr>
</tbody>
</table>

\*Number of registered hospitals
\textsuperscript{†}Number of “outlier” hospitals
Department to find the answer to that question. It is important because health care surveillance methods have not dealt effectively with monitoring multiple units or the overdispersion of health care data (Woodall, 2006).

### 4.3.4 An Alternative Monitoring Scheme

The newspaper report that cardiac patients at the Columbia Presbyterian hospital had the highest postoperative mortality rate of all New York State hospitals highlights a statistical issue with report cards. The judgement that quality of care provided was the worst available was based on the hospital’s RAMR for the year 2001 only. The 2001 report was viewed in isolation with no consideration given to the hospital’s RAMRs in any of the previous years although all of them showed that in-hospital mortality following CABG surgery was not significantly different from the state average.

An alternative to report cards is to publish the mortality trend for individual hospitals or surgeons. For example, a hospital in London (St George’s Hospital) uses the internet to publish control charts to monitor the post-operative mortality rates of the cardiac surgeons who operate there. Such analyses put the variability, consistency, and trends into perspective and avoid dubious institution by institution comparisons. The charts in Figure 4.1 give examples of the use of control charts to monitor the trends in the RAMRs derived from the CSRS database.

The plots in Figure 4.1 (a) chart the RAMRs of the Columbia Presbyterian hospital and the state average OMRs for the years 1994 to 2004. The vertical bars give the published confidence intervals of the RAMRs. The plots show that only the RAMR for 2001 was significantly greater and none were significantly less than the state average OMR. It was suggested in Section 4.3.3 that the true confidence level of the confidence intervals given the New York State report cards is approximately 90%. If such a level of confidence is assumed, then one in ten RAMRs would be expected to be significantly different from the state average because of chance. In the control chart in Figure 4.1 (a), there has only been one out-of-control signal from eleven observations for the Columbia Presbyterian hospital.

The second control chart in Figure 4.1 (b) monitors the RAMRs of the St Joseph’s
FIGURE 4.1: Plots showing the trends of the Risk Adjusted Mortality Rates for CABG surgery and associated 95% confidence intervals of (a) Columbia Presbyterian and (b) St Joseph’s Hospital are given by the solid lines and filled triangles. The time series of dotted lines and open triangles gives the state average OMR for each year from 1994 to 2004.
hospital, which is another New York State cardiac surgical hospital reporting to the CSRS database. The control chart shows that, although the RAMRs of St Joseph’s hospital were less than expected from 1994 to 1998 and no RAMR was above the state average, there was an upward trend from 1997 to 2001, where the RAMRs have increased approximately fivefold from 0.44 to 2.34. This upward trend is an indication that the processes at St Joseph’s hospital are unpredictable and an investigation to establish the cause of the trend should be undertaken. However, such trends are not evident if the RAMRs are published in the form of report cards.

4.4 Simple Methods to Analyse Multi-Institutional Outcome Data for Yorkshire Hospitals

The data used for the analyses in this section were collected between April and October 2003 at eleven hospitals in Yorkshire, England. The expected probability of mortality, the mortality outcome after 30 days of being admitted to hospital, the hospital name, and date of admission were recorded for each of 2,498 patients suffering from acute myocardial infarction (AMI). The hospitals have been de-identified by randomly assigning each a number from one to eleven in place of its name. The coronary care units (CCUs) from which the data were collected are not identical. For example, some CCUs are better equipped than others. Another between-CCU difference is that some have an experienced cardiologist on call 24 hours per day, but others are staffed after-hours by clinicians who are not as highly trained.

The purpose of the analyses described in this section is to detect any hospital mortality rates that are significantly different from the database mortality rate after adjusting for differences in the expected probability of mortality of the hospitals’ patients. The risk score used to risk adjust the patient populations is described in Section 4.4.1. The hospitals with the highest and lowest observed mortality rates are identified and the results of exploratory analysis using notch plots to identify the hospitals with their mean expected probabilities of mortality significantly dif-
ferent from the average of the data set are given in Section 4.4.2. The differences between risk adjusted hospital mortality rates and the data set mortality rate are found using three analytical methods. In Section 4.4.3, the analysis is undertaken with a Poisson model with fixed effects which is appropriate for an assumption that the CCUs are different from one another and, in Section 4.4.4, the data is analysed using a Poisson model with random effects because it is assumed that the CCUs are broadly similar. In Section 4.4.5, the risk adjusted CUSUM is chosen as the analytical tool because the purpose of the analysis is assumed to be prospective sequential monitoring of the processes in each CCU.

### 4.4.1 Risk Adjustment Models for Coronary Care Patients

In a prior study of patient mortality in CCUs in Yorkshire, England, the Evaluation of Methods and Management of Acute Coronary Events (EMMACE) Study Group proposed a relatively simple risk adjustment tool for the expected mortality within 30 days of patients diagnosed with AMI being admitted to hospital (Dorsch et al., 2001). This model used just three explanatory variables to estimate a patient’s probability of mortality 30 days after admission. They were the patient’s age, systolic blood pressure, and heart rate on admission. The advantage of this model is that the data for the explanatory variables were readily available from ambulance or casualty cards. These data were objective and impartial with little room for varying interpretation and manipulation.

Dorsch et al. (2001) found that the predictive power of this model, as measured by the AROC, was 0.79 for their training data and 0.76 for their test data. In fact, they found the AROC of a simple model with age as the covariate was 0.71 for the combined data compared with 0.77 for the three factor model.

Krumholz et al. (1999) compared the AROCs for seven more complex models, which were developed for risk adjustment following AMI, and found the AROCs to vary from 0.71 to 0.78 for training data and 0.70 to 0.78 for validation data. Normand et al. (1996) used logistic regression to formulate a risk model, which used admission characteristics for explanatory variables, to predict patient mortality following hospitalization after AMI. However, their model, which has twenty risk
factors, explained only 27% of the variation in their data. After factors relating to the hospital interventions were added to the model, the unexplained variation reduced slightly from 73% to 67%. They determined these residual percentiles using a modified version of the coefficient of determination (Nagelkerke, 1991) that ranges between 0 and 1.

This brief review of risk adjustment for mortality following admission to a CCU indicates that the performance of the EMMACE model, which has only three risk factors, appears to be comparable to that of models with more risk factors. It supports the conclusion in Chapter 3 that only a limited number of well-chosen risk factors is necessary for the construction of risk adjustment models, provided the model is used within an institution to adjust for the risk of adverse outcomes.

In contrast to its use in a single institution, the application of a risk adjustment model in a process to compare institutions is problematical. As stated in Section 4.2, the purpose of risk adjustment when monitoring outcomes for quality improvement is to remove the variation due to heterogeneity caused by the patient case mix. The underlying assumption is that the remaining variation reflects the quality of care given to the patient and common cause variation (Iezzoni, 1997). However, Normand et al. (1996) concluded that, whilst comparison of mortality rates across hospitals is theoretically possible, such comparative results should be interpreted with the realization that much of the variability for mortality rates cannot be explained by current risk-adjusted mortality models. This review and the literature reviews of trauma scores and cardiac surgical scores in Chapters 2 and 3 suggest that the factors responsible for between institution variation in adverse outcomes are complex and ill-understood. No models which use measures of institutional characteristics to risk adjust the expected probability of adverse patient outcomes were found in the literature.

### 4.4.2 Exploratory Analysis

The unadjusted mortality rates for each of the eleven hospitals in the study are listed in Table 4.2 under the heading Obs Mort. They show that Hospitals 10 and 2 had the lowest death rates of 0.032 and 0.068, respectively, and that Hospitals 7
and 4 had the highest death rates of 0.180 and 0.219, respectively. There is considerable between-unit variability in the observed mortality rates for the CCUs. For example the mortality rate at Hospital 4 was approximately treble the mortality rate of Hospital 10. The mortality rate 0.219 for Hospital 4 was almost twice the mean mortality rate of 0.118 (295 of 2,498 patients) for all subjects in the dataset.

This comparison of the mortality rates in the two hospital assumes that the hospitals and their patient populations are comparable. However, this may not be true. For example, a regional hospital may not have specialist facilities for the treatment of acute myocardial infarction. In such circumstances, seriously ill heart attack victims are stabilized and then transferred to a centre which can provide appropriate specialist treatment. Consequently, it is reasonable that the observed mortality rate of patients admitted to that institution is lower than average because the risk profile of the patient population at that hospital may be different from the risk profile of patients treated in a hospital with specialist facilities. In this section, some exploratory analysis of the predicted probabilities of mortality is undertaken to ascertain if the risk profiles of patients admitted to each hospital are similar.

The mean expected probability of mortality for all patients in the data set is 0.120, which is close to the observed mortality rate. The mean predicted probabilities of mortality for each hospital are listed in Table 4.2. They show less variation than the observed mortality rates, ranging from a minimum of 0.096 for Hospital 9 to a maximum of 0.156 for Hospital 11. However analysis of the distributions of the hospitals' risk scores indicate that there are differences in the populations in each of the hospitals.

The logit transformation of the expected probabilities were found to have an approximately normal distribution and ANOVA analysis (Moore and McCabe, 1993) of the transformed probabilities provided evidence, at the 5% significance level, that there was at least one hospital for which the mean of the transformed probabilities was statistically different from the means of the other hospitals. The notched box plots (McGill et al., 1978) in Figure 4.2 permit comparison of the expected mortality rates of patients from each hospital with the distribution of all patients in the data set. In each of the plots in Figure 4.2, the lightly shaded boxes represent the interquartile range of all the expected mortality rates for all
Table 4.2: Standardized mortality rates (SMR) with 95% Credible Intervals (CI) for acute myocardial infarction patients admitted to 11 Yorkshire hospitals between April and October 2003. For fixed effects, SMR = Obs Mort/Exp Prob and CI was estimated using a standard conjugate Gamma model, as described in Section 4.4.3. For random effects, SMR and CI were estimated using a conjugate Gamma-Poisson hierarchical model, as described in Section 4.4.4. Note that Hospitals 4, 7 and 10 have SMRs that are significantly different from the expected SMR of 1 for the fixed and random effects models.

<table>
<thead>
<tr>
<th>Hosp No</th>
<th>Total Admissions</th>
<th>Obs Mort</th>
<th>Pred Prob</th>
<th>Fixed Effects SMR</th>
<th>95% C.I.</th>
<th>Random Effects SMR</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>181</td>
<td>0.149</td>
<td>0.144</td>
<td>1.04 (0.69, 1.47)</td>
<td>1.04 (0.70, 1.44)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>132</td>
<td>0.068</td>
<td>0.098</td>
<td>0.70 (0.32, 1.22)</td>
<td>0.76 (0.39, 1.24)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>243</td>
<td>0.107</td>
<td>0.131</td>
<td>0.81 (0.53, 1.16)</td>
<td>0.83 (0.56, 1.16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>187</td>
<td>0.219</td>
<td>0.142</td>
<td>1.54 (1.11, 2.05)</td>
<td>1.48 (1.07, 1.96)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>136</td>
<td>0.147</td>
<td>0.107</td>
<td>1.37 (0.84, 2.04)</td>
<td>1.31 (0.83, 1.90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>340</td>
<td>0.100</td>
<td>0.117</td>
<td>0.85 (0.59, 1.16)</td>
<td>0.86 (0.61, 1.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>228</td>
<td>0.180</td>
<td>0.120</td>
<td>1.49 (1.07, 1.98)</td>
<td>1.44 (1.05, 1.91)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>362</td>
<td>0.088</td>
<td>0.111</td>
<td>0.80 (0.54, 1.09)</td>
<td>0.81 (0.57, 1.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>251</td>
<td>0.131</td>
<td>0.096</td>
<td>1.37 (0.94, 1.87)</td>
<td>1.33 (0.93, 1.80)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>249</td>
<td>0.032</td>
<td>0.106</td>
<td>0.30 (0.13, 0.54)</td>
<td>0.38 (0.18, 0.63)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>189</td>
<td>0.127</td>
<td>0.156</td>
<td>0.82 (0.52, 1.17)</td>
<td>0.83 (0.55, 1.18)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Figure 4.2:** Box plots showing the distributions of the patient populations in some Coronary Care Units in north-east England stratified according to the patients’ expected probability of mortality. The light grey boxes represent the interquartile range for all hospitals; the dark grey boxes the interquartile range for each hospital; the whiskers and outliers are for each hospital. The light grey and dark grey notches do not overlap for Hospitals 2, 4, 9, & 11. This is evidence that the median predicted probability of mortality for those hospitals are significantly different, with 95% confidence, from the median probability for all hospitals.
patients and the dark boxes the interquartile range of the expected mortality rates for patients in each hospital. The boxes are notched about the medians of the distributions. The sizes of the notches are chosen such that they are an approximate but conservative 95% confidence test that, if there is no overlap of the notches, the medians of the two distributions are different (McGill et al., 1978). In Figure 4.2, we see evidence that the medians of the predicted risk distribution of the patient population for Hospitals 4 and 11 are significantly higher than the regional median and significantly less for Hospitals 2 and 9.

The exploratory analysis shows Hospital 2 had the second lowest observed mortality rate during the study period, but it also shows that the risk profile for the hospital’s patients was significantly lower than that for all patients in the data set. On the other hand, Hospital 4 had the highest observed mortality rate, but the risk profile of its patients was significantly higher than that for all patients in the data set. Clearly, there is a need for more investigation before coming to any conclusions about the effectiveness of the treatment in any of the CCUs.

### 4.4.3 Analysis with Fixed Effects Model

In the respective analyses in this section and Section 4.4.4, fixed effects and random effects models are used to compute 95% credible intervals for the standardized mortality rates (SMR) in each hospital to enable inferences to be made on the effectiveness of treatment in the CCUs. The SMR is used as a measure of risk adjusted mortality. For this data, it is defined as

\[
SMR = \frac{\sum_j y_j}{\sum_j p_j},
\]

(4.1)

where \( y_j \in \{0,1\} \) takes the value 0 for the event the patient \( j \) survived and 1 for the event the patient \( j \) died within 30 days of admission to hospital, and \( p_j \in (0,1) \) is the expected probability of mortality for patient \( j \) predicted by the EMMACE score.

In the Bayesian context, random effects models are known as hierarchical models. Gelman et al. (1995, Chapter 5) use a study of the effectiveness of cardiac treatments in different hospitals, to illustrate hierarchical models. They state
that, for patients in Hospital $j$ having a survival probability of $\theta_j$, it might be reasonable to expect that the estimates of $\theta_j$s, which represent a sample of hospitals, should be related to one another. That is, the $\theta_j$s are viewed as a sample from a common population distribution. Gelman et al. note that hierarchical modelling attempts to estimate the population distribution from all the (available) data, and thereby to help estimate each $\theta_j$. By contrast, in a fixed effects model, each $\theta_j$ is estimated separately using only the data collected for Hospital $j$. However, a fixed effects model might be preferred on the grounds that the CCUs actually differ and, therefore, the $\theta_j$s should not be viewed as a sample from a common distribution.

In this section, a fixed effects model is used to estimate the uncertainty associated with the SMR for each CCU in the Leeds study. The choice of a fixed effects model may be justified because of differences between the staff and/or equipment in the CCUs described in Section 4.4 and the differences in the risk profiles of the CCU populations found in Section 4.4.2. A fixed effects model for the count of observed mortalities $\sum_j y_j (= Y)$ in each CCU was chosen as

$$Y \sim \text{Poisson}(\lambda S)$$  \hspace{1cm} (4.2)

where $\lambda = \text{SMR}$ and $S = \sum_j p_j$ with each $p_j$ assumed known. If a standard non-informative prior, $p(\lambda) = \lambda^{-1}$, is assumed, then $\lambda|Y$ has the exact posterior distribution (Gelman et al., 1995, Page 48)

$$\lambda|Y \sim \text{Gamma}(Y, S).$$  \hspace{1cm} (4.3)

The SMRs estimated and associated 95% credible intervals are listed in the Fixed Effects columns of Table 4.2. Hospitals 4 and 7 have SMRs that are significantly higher than 1 indicating that the observed mortalities are exceeding the risk-adjusted expected mortalities and the SMR for Hospital 10 is statistically significantly less than expected. The follow up to this statistical analysis should be to establish the reason the hospital have observed mortality rates different from the expected mortality rate. For example, the additional information in Section 4.4 is that some hospitals are better equipped than others to treat AMI. An investigation might find that Hospital 10 is without specialist CCU facilities. Its policy for treatment of patients admitted with a diagnosis of AMI might be to first stabilize
the seriously ill patients and then transfer them to a hospital better equipped to treat AMI. That is, the reason for the low observed mortality rate is that the patients most at risk of dying are moved to another better equipped institution as soon as it is practicable.

### 4.4.4 Analysis with Random Effects Model

For a monitoring scheme that compares the risk adjusted mortality rates of CCUs at different institutions, there is an implicit assumption that each CCU is broadly the same. In such circumstances, a random effects model is the appropriate model. In the method chosen for this analysis, the posterior distributions \( \lambda \mid Y \) were found using BUGS software (MRC Biostatistics Unit, Cambridge, 2005) and a conjugate Gamma-Poisson hierarchical model (Spiegelhalter et al., 2003b, Examples, Volume 1) in which, for each hospital \( i \), where \( i \in \{1, \ldots, 11\} \), the numbers of observed mortalities are distributed

\[
Y_i \sim \text{Poisson}(\lambda_i S_i) \quad (4.4)
\]

and each SMR has the conjugate prior distribution

\[
\lambda_i \sim \text{Gamma}(\alpha, \beta) \quad (4.5)
\]

where the respective priors for the parameters \( \alpha \) and \( \beta \) are the Gamma(1,1) and Gamma(0.1, 0.1) distributions.

The estimates of the SMR with 95\% credible intervals in the Random Effects columns of Table 4.2 show that the SMRs of Hospitals 4 and 7 are significantly higher than expected and that of Hospital 10 is significantly less than expected. The inferences that the SMRs in Hospitals 4 and 7 are higher than expected and that of Hospital 10 is less than expected are the same for both models. However, the results in Table 4.2 indicate that the SMRs found using the random effects model are nearer their mean of 1.01 than the SMRs found using the fixed effects model are to the data set mean.

This shrinkage to the mean is characteristic of hierarchical models. The effect of such models when using comparative schemes to monitor the risk adjusted adverse outcome rate of institutions is to reduce the number identified with a higher or
lower than expected outcome rate. In Section 2.2.3, the potential of Bayesian hierarchical modelling to find fewer “outlier” institutions than modelling with fixed effects was illustrated in a comparative study of nursing homes. The fixed effects model identified 15 of 108 nursing homes as outliers, but the Bayesian hierarchical model only identified two outliers (Berlowitz et al., 2002).

In a quality program based on statistical process control, the next step after finding statistical evidence that the numbers of deaths after treatment for AMI at Hospitals 4 and 7 are greater than expected is to investigate the apparent increase. It is only after the investigation identifies why the processes and procedures in the CCUs are unpredictable that an appropriate program of corrective actions is planned and then implemented.

4.4.5 Analysis with Risk Adjusted CUSUM Schemes

The analyses of mortality outcomes undertaken in the previous sections are retrospective. The issue with retrospective analysis for statistical process control is that any signal that processes are out of control is not timely. The risk adjusted CUSUM chart (Steiner et al., 2000), which is a prospective monitoring tool that allows for varying patient risk of an adverse outcome, was chosen to sequentially monitor the mortality outcomes of the patients admitted to the eleven CCU units in north east England between April and October 2003.

In the examples given by Steiner et al. (2000), \( p_t \) was defined as the expected probability that patient \( t \) would experience an adverse outcome. Their upward risk adjusted CUSUM scheme was tuned to detect a shift in the odds \( p_t/(1 - p_t) \) to the odds \( 2p_t/(1 + p_t) \). That is, the upward CUSUM scheme was tuned to detect a constant shift in the odds ratio from 1 to 2. The downward CUSUM scheme in their example was tuned to detect a shift in the odds ratio from 1 to 1/2.

The tuning used for CUSUM examples in this section followed that used by Steiner et al. (2000). Upward and downward risk adjusted CUSUM charts to monitor the outcomes for Hospitals 4, 7, 10, and 11 are presented in Figure 4.3. The control limits of both the upward and downward CUSUM charts are set at 4 and -4, respectively. This choice of the decision thresholds, which was made
CHAPTER 4: MULTI-INSTITUTIONAL MONITORING

Figure 4.3: Upward and Downward CUSUM Charts monitoring post-admission mortality in the coronary care units of Hospitals 4, 7, 10 and 11. For Hospitals 4 and 7 the Upward CUSUMs have exceeded the decision boundary of 4 signalling shifts in the odds ratio of observed to expected mortality rates from 1 to 2. For Hospitals 10 and 11 the downward CUSUMs have signalled shifts in the odds ratio from 1 to 1/2.

after consideration of the control limits used in the example given by Steiner et al. (2000), is somewhat arbitrary. Usually, decision intervals are chosen so that the average run length (ARL) when the process is predictable is such that the number of false alarms is tolerable, but the ARL when the process is unpredictable is sufficiently short for the monitoring process to be effective.

The upward CUSUM chart monitoring outcomes for Hospital 10 in Figure 4.3 signalled on 23 June 2003 and again on 8 August 2003 that the observed mortality rate of AMI patients admitted to the CCU was higher than expected. There was also a signal from the upward CUSUM chart monitoring outcomes in Hospital 7 that the process had become unpredictable on 6 August 2003. If “on line” mon-
itoring provided by the CUSUM charts had been implemented, the CCU staff in Hospitals 10 and 7 could have responded in June and August, respectively, to the signals that deaths in their hospitals were higher than expected. Their response would have been to find the reason(s) for the signal and, if necessary, take corrective action to stabilize the process. Although the retrospective analyses using fixed and random effects models described in Sections 4.4.3 and 4.4.4, respectively, also indicate that the mortality rates in the CCUs of Hospitals 10 and 7 are greater than expected, the CCU staff would not be aware that processes are out of statistical control until some time after 31 October 2003.

The downward CUSUM chart for Hospital 10 signalled that the mortality rate in the CCU was less than expected on 29 April 2003 and 2 July 2003. Although the out-of-control signals from the downward CUSUM schemes are an indication of better than expected performance, it is necessary that the CCU staff establish the reason for the signals. It was suggested in Section 4.4.3 that the hospital’s policy of transferring high risk patients to other hospitals with specialist facilities for the treatment of AMI is the reason for its lower than expected observed mortality rate. In such circumstances, the EMMACE score should be recalibrated so that the expected probabilities of mortality that it predicts reflect the expected mortality rate in Hospital 10. Otherwise, the upward CUSUM chart is unlikely to signal if there was an upward shift in its observed mortality rate to the rate expected in the other hospitals. Such an observed mortality rate in Hospital 10 would be an indication that its processes are out-of-control.

The SMRs and their 95% Credible Intervals in Table 4.2 do not indicate that the observed mortality rate in the CCU in Hospital 11 was statistically significantly different from the expected rate. However, the downward CUSUM chart for Hospital 11 in Figure 4.3 signalled that the observed mortality rate was less than expected on 12 July 2003. From that date, it indicated that the number of deaths in the CCU was as expected. This chart provides an example that “on line” monitoring of adverse outcomes using risk adjusted CUSUM charts is more informative about and provides more timely warnings of changes in outcome rates than the retrospective analyses given in Sections 4.4.3 and 4.4.4.
4.5 Discussion

The results in Section 4.4 show that each of the three methods of analysing mortality outcomes for patients admitted to CCUs following AMI identified higher than expected risk adjusted mortality rates in hospitals Hospitals 4 and 7. For these data, those CCUs also had the highest observed mortality rate, but medical practitioners assume that outcomes result from a complex mix of factors and that risk adjustment removes one source of variation from the mix leaving residual differences to reflect quality of care (Iezzoni, 1997). For example, although the observed mortality rate for Hospital 4 is highest, Table 4.2 shows that its mean expected probability of mortality was only exceeded by Hospital 11.

The report cards for adult cardiac surgery in New York State hospitals do provide risk adjusted mortality rates, but there are many criticisms of their publication. Examples of these criticisms, given in Section 4.3.2, include a marked variation of the rankings of hospitals from year to year and general scepticism about the quality of the data used to RAMRs. One way to mitigate the uncertainty in the estimated RAMRs is their publication using trend analysis, such as that proposed in Section 4.3.4. It appears, however, that medical practitioners are opposed to any publication of quality measures of the services that provide. The debate on the use of report cards, which was reviewed in Sections 4.2 and 4.3.2, suggests the reason is they doubt that the measures used give a true indication of the quality of care provided.

The argument that publication of performance indicators for hospitals or medical practitioners does not give a true indication of the quality of care has been supported by the statistical community. For example, Goldstein and Spiegelhalter (1996) discussed the limitations of league tables and urged caution in their use. However, Werner and Asch (2005) have suggested that this is not the only issue. They raised the issue that publication of performance indicators allows the public to hold health care providers accountable for quality of care they deliver. For example, in Queensland, the performance indicators of the state’s public hospitals were made public after a scandal about the care provided by a surgeon at the Bundaberg Base Hospital. A Commission of Inquiry (Davies, 2005) was raised
and there was political pressure for the publication of the “secret” performance indicators when their existence became public knowledge (Cole, 2005).

It is clear that publication of health care performance indicators for Queensland public hospitals was inevitable, but it appears that the Queensland public hospitals’ performance report was issued without input from professional bodies such as the Australian Medical Association or comment from any statisticians within the state. If this is so, it is a mistake. Tu et al. (2001) advocate that clinicians be involved in report card development. Otherwise, their production may be dominated by individuals with less clinical insight, and misleading conclusions are more likely.

The Royal College of Physicians and other professional bodies provide an example of how clinicians may influence publications on healthcare providers with their input into the issue of report cards on the treatment of heart attack patients in hospitals in England and Wales. In 1995, the sensitivity of mortality outcomes of patients admitted to hospital for treatment of acute myocardial infarction to changes in the quality of care provided was questioned (Davies and Crombie, 1995; Mant and Hicks, 1995). Changes in clinical processes used to treat heart attack victims were considered better indicators of the quality of care provided and, accordingly, a scheme to monitor the clinical process was proposed. By 2000, a group working under the aegis of the Royal College of Physicians and British Cardiac Society proposed a national audit where data were collected by clinicians employed by the hospitals and the audit process satisfied the requirements of the National Service Framework (Birkhead, 2000). The proposal was approved and funded by the Department of Health. The Myocardial Infarction National Audit Project (MINAP) was implemented and has published five annual reports on how the National Health Service (NHS) manages heart attacks (Royal College of Physicians, 2002–2006). Each includes a report card giving information about clinical process measures, such as the proportion of a hospital’s patients having thrombolytic treatment within 30 minutes of admission to hospital, rather than a report card on hospitals’ mortality rates.

According to the 2006 report, MINAP allows hospitals and ambulances to measure their performance against national standards and targets for the care of heart
attack patients. MINAP states the continuous collection and publication of data enables comparison of performance across the NHS and monitors the improvement in care for people who have suffered a heart attack. The data to monitor if the treatment NHS hospitals provide meets national standards are collected by health care workers. It appears MINAP report cards are being used in a positive manner.
Chapter 5

Performance of Risk Adjusted CUSUM Schemes: Expected Outcome Probability, $p_t$, Assumed Known

The Cumulative Sum (CUSUM) control chart (Page, 1954) is an effective monitoring tool for detecting small shifts (Montgomery, 2001) in a parameter of interest $\theta$. Suppose $\theta = \theta_0$ when the process is in statistical control and $\theta = \theta_1$ when the process is out of statistical control. If $\theta_0 < \theta_1$, CUSUM schemes take the upward form

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

(5.1)

where $C_t$ is the CUSUM value at observation $t$ and $W_t = \log \left\{ f(y, \theta_1)/f(y, \theta_0) \right\}$, the log likelihood ratio, is known as the CUSUM Weight. An alarm is signalled when $C_t \geq h$, some predetermined decision threshold, and monitoring may commence at any $0 \leq C_0 < h$ although the discussion in this chapter is limited to $C_0 = 0$. For $\theta_0 > \theta_1$ the downward form of the CUSUM is

$$C_t = \min(0, C_{t-1} - W_t)$$

(5.2)

with a signal when $C_t \leq -h$. The formulations in Equations (5.1) and (5.2) are known as the decision interval form of the CUSUM (Hawkins and Olwell, 1998).

In this chapter, the performance of the risk adjusted CUSUM scheme (Steiner...
et al., 2000) is assessed. It was designed to monitor adverse outcomes following medical procedures where the expected probabilities of the outcomes change because of the heterogeneity of the patient population. The outcomes $Y$ take the value 0 or 1 and are modelled with the distribution $Y \sim \text{Bernoulli}(\pi)$, where $\pi$ is the probability of the event $Y = 1$ and, in the discussion that follows, $\pi$ will be the parameter of interest.

The discussion commences in Section 5.1 with a description of a CUSUM scheme which is used in the industrial context to monitor the proportion of manufactured items that do not conform to specification. A review of the literature on its adaption for monitoring mortality outcomes after adult cardiac surgery follows. In Section 5.2 some methods of assessing the performance of CUSUM schemes are introduced and reasons for choosing the ARL as the assessment tool are given. The risk adjusted CUSUM scheme is described and the method and results of a simulation study to find the ARLs as the risk profile of the population varies are presented in Section 5.3. The results show that, for low outcome rates, both the in-control and out-of-control ARLs of the upward risk adjusted CUSUM increase rapidly as the average expected mortality rate of the patient population decreases. These results motivated the examination of the properties of the risk adjusted CUSUM scheme given in Section 5.4. The changes in the risk adjusted CUSUM weight $W_t$, the mean $\text{E}(W_t)$ and the variance $\text{Var}(W_t)$ of the weight, and the relationship between its in-control parameter value, $\pi = \pi_0$, and out-of-control value, $\pi = \pi_1$, as the expected probability of mortality, $p_t$, predicted by the risk score, are discussed. One hypothesis arising from the analysis in Section 5.4 is that the ARL of the risk adjusted CUSUM is stable if the step size from the parameter value for an in-control process to the parameter value for an out-of-control process, $|\pi_1 - \pi_0|$, is not less than 0.05. The results of a simulation study to test this hypothesis are presented in Section 5.5. There is a digression in Section 5.6 where the results of an exploratory study to compare and contrast the performance of the binomial CUSUM scheme (Hawkins and Olwell, 1998) with that of the risk adjusted CUSUM scheme are presented. Section 5.7 contains some discussion and a conclusion that the performance of the risk adjusted CUSUM scheme is improved if it is tuned so that there is always a minimum step between the parameter value $\pi_0$
for the process in-control and parameter value $\pi_1$ for the process out-of-control.

\section*{5.1 Review of CUSUM Schemes}

\subsection*{5.1.1 Industrial CUSUM Schemes}

The CUSUM schemes designed to monitor situations for which each observation can be one of two classes use the binomial distribution as a statistical model. Hawkins and Olwell (1998, Chapter 5.6) state that the classes are traditionally those that conform and those that do not conform. In their scheme, the number of nonconforming items in a sample of size $m$ follows a binomial distribution with probability of $\pi$ subject to the following constraints:

1. the sample size $m$ is fixed in advance;

2. each item is conforming or nonconforming;

3. the probability, $\pi$, that an item is nonconforming is constant; and

4. the different items are statistically independent of each other.

Then the probability $f(y_t, \pi)$ that there are $y_t$ nonconforming items in the $t^{th}$ sample, where $y_t \in \{0,1,\ldots,m\}$ and $t = 1,2,\ldots$, is given by

$$f(y_t, \pi) = \binom{m}{y_t} \pi^y_t (1-\pi)^{m-y_t}, \quad \pi \in (0,1) \quad (5.3)$$

The CUSUM, tuned to detect a shift from an in-control parameter value $\pi_0$ to an out-of-control value $\pi_1$, has the weight $W_t$ given by the log-probability ratio

$$W_t = \log \frac{\binom{m}{y_t} \pi_1^{y_t} (1-\pi_1)^{m-y_t}}{\binom{m}{y_t} \pi_0^{y_t} (1-\pi_0)^{m-y_t}}$$

$$= \gamma \left\{ y_t + m \frac{\Delta}{\gamma} \right\} \quad (5.4)$$

where $\gamma = \log \left\{ \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \right\}$ is the scaling factor and $\Delta = \log \left\{ (1-\pi_1)/(1-\pi_0) \right\}$ is the tuning factor.

Hawkins and Olwell (1998) propose a binomial CUSUM scheme first derived by Johnson and Leone (1962) in which the CUSUM weight is defined as

$$W_t = y_t + m \frac{\Delta}{\gamma} \quad (5.5)$$
CHAPTER 5: RA-CUSUM—PREDICTION KNOWN

\[
C_t = \max \left\{ 0, C_{t-1} + \left( y_t + m \Delta \gamma \right) \right\} \quad (5.6)
\]

and, for \( \pi_0 > \pi_1 \), the downward scheme

\[
C_t = \min \left\{ 0, C_{t-1} + \left( y_t + m \Delta \gamma \right) \right\}. \quad (5.7)
\]

For this scheme \( \Delta \gamma < 0 \) if \( 0 < \pi_0 < \pi_1 < 1 \) or \( 0 < \pi_1 < \pi_0 < 1 \).

Figure 5.1 provides an example of an industrial CUSUM chart tuned to detect relatively small upward or downward shifts in the proportion of tins packed with frozen orange juice concentrate that could possibly leak. Initial analysis (Montgomery, 2001, Chapter 6.1) has established that the expected fraction of nonconforming cans is 11.1%. The charts, which use the CUSUM designs given in Equations (5.6) and (5.7), are tuned to detect upward and downward shifts in the nonconforming fractions in random samples of size 50 from 11.1% up to 17.7% and down to 4.5%, respectively. The step size of 6.6%, which is a shift of 1.5 \( \sigma \) where \( \sigma \) is the standard error, was chosen because Montgomery (2001, Chapter 8) suggests CUSUM schemes are effective in detecting small shifts in the process, say 1.5 \( \sigma \) or less. The decision intervals of 6.5 and -4.5 were chosen because the average run length to a signal when the process is out-of-control (ARL\(_1\)) is approximately four for both upward and downward CUSUMs and the average run length to a signal when the process is in-control (ARL\(_0\)) is 170 for the upward CUSUM and 369 for the downward CUSUM. The CUSUM performance measures ARL\(_0\) and ARL\(_1\) are discussed in detail in Section 5.2.

For the CUSUM scheme in Figure 5.1, the decision rule is to take action if there is a shift of 6.6% in the proportion of nonconforming tins. After the 78th sample, the CUSUM signalled the process was out-of-control and the next action was to investigate for any assignable causes and take corrective action if necessary. Montgomery (2001) makes the point that the fraction of nonconforming items of 11.1% for an in-control process is too high but it will require management intervention to make process changes to reduce common cause variation responsible for the excessive proportion of nonconforming items.

Reynolds, Jr and Stoumbos (1999) propose a Bernoulli CUSUM scheme to monitor for an increase in the proportion of nonconforming items where the re-
Figure 5.1: Illustrative example of upward and downward binomial CUSUM schemes tuned to detect relatively small shifts of 1.5 standard deviations in the proportion of nonconforming manufactured items. Data Source: Montgomery (2001, Chapter 6.1)
result of the inspection of an individual item are available immediately after its inspection. They suggest that such a scheme the would be useful in situations where the in-control value $\pi_0$ is low and note that a decision regarding the state of the process may be made following each inspection. The design of the Bernoulli CUSUM scheme, which is similar to that of the upward binomial CUSUM scheme in Equation (5.6) except that $m = 1$, is given as

$$C_t = \max (0, C_{t-1}) + \left( y_t + \frac{\Delta}{\gamma} \right).$$

(5.8)

Such a design allows negative values to be plotted on the CUSUM chart. Reynolds, Jr and Stoumbos also provide two new methods of approximating the ARL$_{0}$ and ARL$_{1}$ of a Bernoulli CUSUM scheme designed to detect a shift from $\pi_0$ to $\pi_1$ of the probability of a non-conforming item and with the decision threshold set to $h$. In the first, they compute the ARLs by explicitly solving the linear equations given by Page (1954, Section 5.1) and, in the second, their evaluation of the ARLs is based on using approximations developed by Wald (1947) and diffusion theory corrections to these approximations that extend the work of Siegmund (1979, 1985).

### 5.1.2 CUSUM Schemes to Monitor Mortality Outcomes

The success of statistical quality control in the industrial context motivated its introduction for paediatric cardiac surgery. For example, Dr de Leval, a paediatric cardiac surgeon, experienced a cluster of deaths following neonatal arterial switch operations. He sought retraining with a surgeon known for a very low mortality rate for this operation. De Leval et al. (1994) provide statistical analyses of data collected prior to, during, and after the cluster of deaths. The analyses included two CUSUM control charts based on the SPRT (Wald, 1947): one to plot the mortality outcomes trend and the other to plot mortality or the need to return the patient to cardiopulmonary bypass after a trial period of weaning. Each chart had two pairs of decision thresholds. De Leval et al. (1994) stated that the first decision boundary was an “alert” line and the second was an “alarm” line corresponding to levels of statistical significance of 0.2 and 0.05, respectively. They concluded that one could be 80% or 95% sure that the acceptable failure rate had been
transgressed if the CUSUM crossed above the alert line or alarm line, respectively. This CUSUM scheme may be written in the upward decision interval form (Steiner et al., 1999) as

\[ C_t = \max \left\{ 0, C_{t-1} + \gamma \left( y_t + \frac{\Delta}{\gamma} \right) \right\}, \quad (5.9) \]

where \( y_t \) takes the value 0 for the event that patient \( t \) survives or 1 for the event that patient \( t \) dies. Similarly the two lower boundaries provide reassurance.

Risk factor analysis identified infants with the origin of the circumflex coronary artery from (aortic) sinus 2 at higher risk of mortality following the arterial switch operation, but an improvement in surgical technique reduced this risk (De Leval et al., 1994). There were no other risk factors clearly identified so De Leval et al. (1994) assumed, a priori, each infant has the same probability of dying after the operation. Therefore, the fixed rate CUSUM scheme given by Equation (5.9), in which the in-control and out-of-control parameter values, \( \{\pi_0, \pi_1\} \), are fixed, is a suitable tool to monitor mortality rates following arterial switch operations.

In contrast to the industrial schemes given in Equations (5.6) and (5.7), the CUSUM in Equation (5.9) retains the scaling factor \( \gamma \) and, therefore, the weight \( W_t \) is the log likelihood ratio where the binary observation \( Y_t \) is assumed independent with distribution \( Y_t \sim \text{Bernoulli}(\pi) \).

The CUSUM scheme used by De Leval et al. (1994) to monitor deaths following neonatal arterial switch operations is tuned to detect a shift in the mortality rate from 0.02 to 0.05. Substituting into Equation (5.9) gives \( \gamma = 0.947 \) and \( \Delta = -0.031 \) and weights

\[ W_t = \begin{cases} 
0.916 & \text{if patient } t \text{ dies}, \\
-0.031 & \text{if patient } t \text{ survives},
\end{cases} \quad \text{(Steiner et al., 1999).} \]

In many medical contexts, there is considerable variation (heterogeneity) in the characteristics of the people under study (Steiner et al., 2000). The risk adjusted CUSUM scheme (Steiner et al., 2000) allows for the heterogeneity by relaxing Constraint 3 in Section 5.1.1 that the probability, \( \pi \), of an outcome is constant. It may be used to monitor populations with varying risk profiles because, for each increment, the weight \( W_t \) depends on the expected outcome rate \( p_t \) of patient \( t \). In their example, Steiner et al. (2000) compare the performance of a consultant
and a trainee cardiac surgeon, who both undertook cardiac surgical operations at a single surgical centre, by monitoring 30-day mortality following CABG surgery. The average expected probability of mortality of the patients treated by the consultant is greater than that of the total population treated at the surgical centre, but mean expected probability of mortality of the trainee’s patients is less than that of the total patient population. When the fixed rate CUSUM scheme (Equation 5.9) is used the upward CUSUM signals that the consultant’s performance is out-of-control and the downward CUSUM signals an improvement in the trainee’s performance but upward and downward CUSUMs using the risk adjusted scheme indicated the performances of both surgeons were in-control. The authors conclude that the sensitivity of the risk adjusted scheme, which allows for a changing mix of patient characteristics, can be set so that false alarms do not happen very frequently, but substantial changes in the failure rate are quickly detected.

The risk adjusted CUSUM scheme appears to be an appropriate tool to monitor any outcome event for which the predicted probability of the outcome varies. The upward and downward charts in Figure 5.2 monitor the death rate of patients admitted to the ICU at the Princess Alexandra Hospital, Brisbane from January 1995 to December 1997. The probability of mortality of each patient was predicted on admission using the APACHE III (Knaus et al., 1991) risk score. The decision intervals of 4.5 and -4.0 are those used by Steiner et al. (2000). Using the simulation approach described in Section 5.3, the candidate found that the approximate ARLs are 5,379 and 3,771, respectively, and the corresponding 95% confidence intervals are (5,346, 5,412) and (3,719, 3,822). The signal from the upward chart in Figure 5.2(a) on 29 June 1995 suggests deaths within the ICU were higher than expected. The unit instituted a quality assurance program in which the number of full time medical staff increased and there was a comprehensive review of clinical practices and procedures (Cook et al., 2002). In Figure 5.2(b), the signal from the downward CUSUM on 22 February 1997 provides some evidence that the improvements implemented by ICU staff resulted in a reduction in patient mortality rate.
Figure 5.2: (a) Plot for an upward risk adjusted CUSUM scheme tuned to signal that the process in an intensive care unit is out of statistical control if the odds ratio of the observed to expected mortality outcomes doubles, $R_A = 2$, and (b) plot for a downward risk adjusted CUSUM scheme tuned to signal if the odds ratio of the observed to expected mortality halves, $R_A = 1/2$. See Equation (5.11) in Section 5.3.2 for the definition of $R_A$.

Data Source: Cook et al. (2003)
5.2 Review of Methods to Assess CUSUM Performance

When considering the problem of monitoring continuous industrial processes, Page (1954) notes that the quality of output may be assessed by some measurable characteristic. This measure has a statistical distribution with some parameter of interest $\theta$, which Page calls a quality number, to give an indication of quality. An issue with continuous inspection schemes is that, in all cases of practical interest, there is a probability of one that some point will eventually fall outside the control limits and action will then be taken although there is no change in the quality number $\theta$. Page chooses the ARL function of a process inspection scheme as the measure of the effectiveness of control chart schemes citing Aroian and Levene (1950) who suggest several measures for assessing the effectiveness of control charts. When the quality number $\theta$ remains constant Page defines the ARL as the expected number of articles sampled before action is taken. He noted that the ARL is a function of $\theta$. When the quality of output is satisfactory, $\text{ARL}_0$ is a measure of the expense incurred by the scheme when it gives false alarms (analogous to Type I errors in hypothesis testing) and, for poor quality, $\text{ARL}_1$ measures the delay and thus the amount of scrap produced before rectifying action is taken (Type II errors).

Hawkins and Olwell (1998) discuss runs, run lengths and ARLs for CUSUM schemes giving the following definitions.

- A run is the sequence of CUSUM values from the starting point $C_0$ to the CUSUM crossing the decision boundary signalling the process is out of statistical control.

- A run length is the number of observations in a run. The run length is a random variable, having a mean, variance, and a distribution.

- Average run length (ARL) is the name given to the mean of the run length distribution.
As Page (1954) does, Hawkins and Olwell (1998) see the means of the run length distributions, $ARL_0$ and $ARL_1$, as summaries analogous to the probabilities of Type I and Type II errors in classical hypothesis testing, but they warn that the measures are less than perfect because the run length has approximately a geometric distribution and, thus, the run length distribution is skewed to the right and has a standard deviation approximately equal to the ARL. Nevertheless, they say the ARL is an easily interpreted, well defined measure and is the standard measure of performance for CUSUM schemes.

The discussion by Page and Hawkins and Olwell referred to above is in the context of monitoring industrial production. For “on line” or prospective monitoring of medical processes, Sonesson and Bock (2003) suggest surveillance schemes with observations $X = \{X_t : t = 1, 2, \ldots\}$ be assessed using one or more of the following:

- $ARL_0$,
- $ARL_1$,
- Probability of a false alarm,
- Conditional Expected Delay ($CED(t)$),
- Expected Delay ($ED_\tau$) where $\tau$ is the unknown time of change from an in-control to an out-of-control process,
- Probability of successful detection when only a delay time of $d$ after the process change may be tolerated,
- Predictive value which is the conditional probability that the process is out of control when an alarm is signalled.

Sonesson and Bock (2003) denote the respective in-control and out-of-control states as $D(s)$ and $C(s)$, where $s = 1, 2, \ldots$, is the decision time. At each decision time $s$ the observations $X_s = \{X(t) : t \leq s\}$ give an indication that the process is in state $C(s)$ if $X_s \in A(s)$, where $A(s)$ are the alarm sets. An alarm is usually
signalled using an alarm function \( p(X_s) \) and a decision rule \( g(s) \), such as a control limit, so that the time of alarm \( t_A \) is defined as

\[
t_A = \min \{ s : p(X_s) > g(s) \}
\]

The case studied by Sonesson and Bock (2003) and Frisén (2003) is when \( D(s) = \{ s < \tau \} \) and \( C(s) = \{ s \geq \tau \} \), but they note other definitions of in-control and out-of-control states are possible.

For the risk adjusted CUSUM scheme a change in the single parameter of interest in the distribution of \( X \) is considered. Each of the performance measures listed above may now be written in terms of the time of alarm \( t_A \), the change point \( \tau \), and the other variables defined in the previous paragraph.

ARL_0, the mean of the distribution of run lengths to false alarms, is denoted by \( E(t_A | \tau = \infty) \). The probability of a false alarm, \( \Pr(t_A < \tau) \), is another measure of false alarms and is given by

\[
\Pr(t_A < \tau) = \sum_{t=1}^{\infty} \Pr(t_A < \tau | \tau = t) \Pr(\tau = t),
\]

where both \( t_A \) and \( \tau \) are random. It is necessary to know the distribution \( \Pr(\tau = t) \) of the change point \( \tau \) before computing \( \Pr(t_A < \tau) \). If the intensity, defined by Frisén (2003) as the hazard \( \nu_t = \Pr(\tau = t | \tau \geq t) \), is constant. Sonesson and Bock (2003) find that a suitable assumption is that the distribution of \( \tau \) is geometric, provided that the intensity of a shift is constant for each time point.

ARL_1, conditional expected delay, and expected delay are measures of performance when the process is out-of-control. \( ARL_1 = E(t_A | \tau = 1) \), which implies the process is out-of-control when monitoring commences, is used in the literature on quality control. Sonesson and Bock (2003) suggest other measures are more appropriate in a public health situation. They define the conditional expected delay, which is the average delay time for an alarm when \( \tau = t \), as

\[
CED(t) = E(t_A - \tau | t_A \geq \tau = t),
\]

and the expected delay, which is a weighted average of the delay before an alarm and for which a distribution for \( \tau \) is assumed, as

\[
ED_\tau = \sum_{t=1}^{\infty} CED(t) \Pr(t_A \geq t) \Pr(\tau = t)
\]
In some processes, only a limited delay $d$ between the change $\tau$ and its detection may be tolerated. The examples given were surveillance for the outbreak of an infectious disease and monitoring foetal heart rate during labour to detect any abnormality. For this class of processes the probability of successful detection $\text{PSD}(d, t)$, defined as

$$\text{PSD}(d, t) = \Pr(t_A - \tau \leq d \mid t_A \geq \tau = t),$$

may be considered. This measure of effectiveness is not appropriate for risk adjusted CUSUM schemes because they are used to monitor for increases or decreases in outcome rates for patients who have completed some medical procedure.

The final measure of effectiveness suggested by Frisén (2003) and Sonesson and Bock (2003) is the predictive value of alarm $\text{PV}(t)$ defined as

$$\text{PV}(t) = \Pr(C(t) \mid t_A = t).$$

The value of $\text{PV}(t)$ gives information on how much trust to put in an alarm occurring after observation $X_t$ and, therefore, allows a choice from a number of possible actions, including no action, to be taken following a signal from the monitoring system. In the philosophy of statistical process control used for the research undertaken in this thesis, monitoring schemes are designed so that the number of false alarms is tolerable but time to a true alarm will be sufficiently short to prevent excessive damage or waste because the process is out-of-control. The single response to an alarm is to investigate and establish its cause. There is no reason to consider the predictive value of an alarm.

For a risk adjusted CUSUM scheme, the performance measures that should be considered are $\text{ARL}_0$ and $\Pr(t_A < \tau)$ for assessing false alarms and $\text{ARL}_1$, $\text{CED}(t)$, and $\text{ED}_\tau$ for assessing the power of the scheme. As noted at the beginning of this section, $\text{ARL}_0$ and $\text{ARL}_1$ have been used to assess the performance of CUSUM schemes based on SPRT since their inception.

Frisén (2003) is critical of the use of the average run lengths as assessment tools because $\text{ARL}_1$ measures the expected $t_A$ for $\tau = 1$ only. To illustrate the shortcomings of the $\text{ARL}_1$ she proposes two surveillance schemes which minimize the $\text{ARL}_1$ for fixed $\text{ARL}_0$. The first is an artificial scheme she calls the Two-Point...
method. She introduces it by proposing that there exists values \( c_s \) such that a surveillance system with alarm at

\[
t_A = \min \left\{ s : \sum_{t=1}^{s} X(t) > c_s \right\}
\]

gives the minimal ARL\(_1\) for fixed ARL\(_0\). She then constructs the Two-Point method with a fixed ARL\(_0\), denoted by \( A \), and the standard normal distribution function \( \Phi \).

The method has alarm limits \( c_1 = L, c_i = \infty, \) for \( i = 2, 3, \ldots, k - 1, \) and \( c_k = -\infty, \) where \( k = \left[ A - \Phi(-L) \right] / \Phi(L) \) and \( L \) is restricted to those values which make \( k \) an integer. She notes that the Two-Point method fulfils the criterion of minimal ARL\(_1\) by having ARL\(_1\) arbitrarily close to the minimal value, one, for fixed ARL\(_0\).

Although the scheme has minimal ARL\(_1\), it has very bad properties because an alarm at time 1 or time \( k \) will depend only on the first observation.

Next, Frisén (2003) gives a more reasonable method which minimizes the ARL\(_1\) for fixed false alarm probability. She proposes a surveillance system with alarm at

\[
t_A = \min \left\{ s : \sum_{t=1}^{s} X(t) > L + s\mu/2 \right\},
\]

where \( L \) is a constant, which gives minimal ARL\(_1\) in the class of methods with the same false alarm probability \( \Pr(t_A < \tau) \). Finally, she criticizes the ARL\(_1\) measure from the viewpoint of optimal decision theory noting it is hard to motivate a cost function with no cost for a delay of an alarm when \( \tau > 1 \).

Nevertheless, Frisén (2003) acknowledges the maximal value of CED\((i)\) is equal to CED\((1)\) and with a minimax perspective this can be motivation for the use of ARL\(_1\) because CED\((1) = ARL_1 - 1\). A typical relationship between ARL\(_1\) and ED\(_\tau\) is ARL\(_1 = ED_1 + 1\) and ED\(_\tau\) tends to zero as \( \tau \) increases. For most schemes CED\((t)\) converges to a constant value sometimes called the steady state ARL.

For CUSUM schemes based on SPRT, the values of the in-control and out-of-control ARLs converge to steady state ARLs which are smaller than the ARLs found using the starting point \( C_0 \), but the relative rankings of different CUSUMs are generally not greatly affected by the assumption about the starting point (Hawkins and Olwell, 1998). Moustakides (1986) shows that, for the worst possible values of \( \tau \) and \( C_{\tau - 1} \), the delay for detecting a process change which has
caused a shift from the in-control parameter value to the out-of-control parameter value is minimal for the CUSUM scheme. Ritov (1990) considers a loss function which depends on $\tau$ and $t_A$ besides $t_A - \tau$. The worst possible distribution $P(\tau = s + 1 | \tau > s; X_s)$ is assumed for each time $s$. Under this assumption of a worst possible distribution (based on earlier observations), the CUSUM minimizes the loss function. That is, when compared to other surveillance schemes, the CUSUM is optimal in this sense and the optimality is independent of the time of a process change. In a recent discussion paper, Woodall (2006) recommends that CUSUM charts used for medical monitoring should be designed based on run-length performance measures such as ARL. It appears the comparison of ARL$_{1s}$ for fixed ARL$_{0s}$ is a satisfactory method for assessing the effectiveness of CUSUM schemes considered in this thesis and, it is suggested, the ARL measure is more readily understood by hospital staff.

5.3 Assessment of the Performance of Risk Adjusted CUSUM Schemes Using the Average Run Length

In this section, the performance of the risk adjusted CUSUM scheme proposed by Steiner et al. (2000) is assessed. The ARL$_{0s}$s and ARL$_{1s}$s of risk adjusted CUSUM schemes, that are monitoring the number of adverse outcomes for patient populations with varying risk profiles, are approximated using simulation and used to assess the performance of each scheme.

The sources and distributions of the patient populations are given in Section 5.3.1. A formal definition of the risk adjusted CUSUM scheme as proposed by Steiner et al. (2000) is given in Section 5.3.2 and, in Section 5.3.3, the method used to carry out the simulations is outlined and the results of the simulation studies on the Steiner CUSUM scheme are presented and discussed.
5.3.1 Distributions of Patient Populations

The risk adjusted CUSUM scheme (Steiner et al., 2000) was initially proposed to monitor individual cardiac surgical performance. The risk of patient mortality following cardiac surgery may be assessed using one of the risk scores discussed in Chapter 3. Two of the scores used to assess adult patients proposing to undergo CABG surgery are the Parsonnet score (Parsonnet et al., 1989) and the EuroSCORE (Nashef et al., 1999). Patient populations may be categorized using one or another of these scores. Figure 5.3(c), taken from Crayford (2000, Figure 1, “Other Consultant”)

The distributions in Figures 5.3(a) and (b) were manufactured by right truncating and rescaling Distribution (c) so that the respective patient populations are classified by Parsonnet scores 0 to 7 and 0 to 11, only. Distribution (a) represents the 60% of the population given in Distribution (c) at lowest risk of post operative mortality and Distribution (b) represents the 80% of the population at lowest risk. A less experienced surgeon might be expected to treat populations (a) and (b). The distributions of the patient populations that a consultant surgeon might treat are given in Figures 5.3(d) and (e). They were manufactured by left truncating and rescaling Distribution (c) so that Distribution (d) represents the 80% of the population at highest risk and Distribution (e) the 60% at highest risk. The patient populations in Distributions (d) and (e) are classified by Parsonnet scores 1 to 47 and 3 to 47, respectively.

The distribution in Figure 5.4(c), which is for all cardiac surgical patients undergoing CABG surgery for the first time between April 1999 and March 2002 in all National Health Service centres in the north west of England (Bridgewater et al., 2003), classifies the patient population by EuroSCOREs 0 to 19. The distributions in Figures 5.4(a), (b), (d), and (e) were synthesized by applying the respective transformations $2/((\sqrt{x^2} + 1)$, $2/(x + 1)$, $0.75 + 0.5x$, and $0.5 + 0.1x$, where $x$ is the EuroSCORE, and rescaling. In this way, patient populations with

---

1The internet site which published Crayfords’s report is no longer active but a copy of the report is available from the author.
Figure 5.3: Distribution of cardiac surgical populations where the risk of mortality is classified by Parsonnet score. The population in Chart (c) is as given by Crayford (2000, Figure 1, “Other Consultant”), the distributions in Charts (a) and (b) represent the 60% and 80%, respectively, of the population given in Chart (c) at lowest risk and the distributions in Charts (d) and (e) represent the 80% and 60%, respectively, of the population at highest risk.
varying risk profiles are manufactured, but the number of risk categories is constant for each of the patient populations in Figure 5.4. The proportion of high risk patients in the populations given in Figures 5.4(a) and 5.4(b) compared to that in Figure 5.4(c) is reduced. The use of transformations to reduce the overall risk of the cardiac surgical population models possible case mixes for an inexperienced surgeon who, on occasion, may undertake surgery on patients with the highest risk. The respective proportions of low and high risk patients are higher and lower than for the patient population. In contrast, a consultant surgeon might treat a case mix, such as one of those given in Figures 5.4(d) or 5.4(e), where the proportions of low risk and high risk patients are lower and higher, respectively, than for the patient population.

Steiner et al. (2000) found that the simple logistic regression model

\[ \logit(p_t) = \alpha + \beta X_t, \]  

(5.10)

where \( p_t \) is the expected probability of mortality and \( X_t \) is the risk score for patient \( t \), for \( t = 1, 2, \ldots \), is an appropriate model to estimate the expected probability of mortality for each patient. Their parameter estimates \((\alpha, \beta) = (-3.68, 0.077)\) have been used when patient risk has been assessed using the Parsonnet score. Bridgewater et al. (2003) provided information on the patient deaths which, when fitted to the model in Equation (5.10) with the EuroSCOREs as the explanatory variable, gave estimated parameter values of \((\alpha, \beta) = (-5.565, 0.340)\).

Using the appropriate parameter values, the average expected mortality rates of the respective patient populations in Figures 5.4 and 5.3 were 0.007, 0.009, 0.017, 0.020, and 0.024 for those categorized by EuroSCORE and 0.030, 0.034, 0.046, 0.051, and 0.058 for those categorized by Parsonnet score.

A population of intensive care patients was manufactured using a data set of 3,398 patients admitted to the ICU at the Princess Alexandra Hospital, Brisbane, between 1 January 1995 and 31 December 1997 to simulate the expected probabilities of mortality of admissions to an ICU. The APACHE III (Knaus et al., 1991) risk score was used to predict the expected probabilities of mortality, which have a mean of 0.160 for the data set.

Each patient population described in this section was used in the simulation
Figure 5.4: Distribution of cardiac surgical populations where the risk of mortality is classified by EuroSCORE. The population in Chart (c) is derived using data published by Bridgewater et al. (2003), the distributions in Charts (a) and (b) approximate the 60% and 80%, respectively, of the population given in Chart (c) at lowest risk and the distributions in Charts (d) and (e) approximate the 80% and 60%, respectively, of the population at highest risk.
studies to estimate $ARL_0$ and $ARL_1$ for both a Steiner risk adjusted CUSUM scheme, described in Section 5.3.2, tuned to detect a doubling of the expected odds ratio and a minimum effect risk adjusted CUSUM scheme, described in Section 5.5. That is, simulation was used to approximate the $ARL_0$s and $ARL_1$s of CUSUM schemes which were monitoring the mortality outcomes of each of the five cardiac surgical patient populations risk assessed by the EuroSCORE, the other five cardiac surgical populations risk assessed using the Parsonnet score, and the ICU patient population risk assessed by the APACHE III score.

5.3.2 Definition of the Risk Adjusted CUSUM scheme

A CUSUM scheme may be thought of as sequence of tests of the null hypothesis $H_0: \pi = \pi_0$ against the alternative $H_1: \pi = \pi_1$. In an upward scheme, testing continues until $C_t \geq h$ indicates $\pi = \pi_1$. Page (1954) describes an equivalent scheme as a sequence of sequential probability ratio tests (Wald, 1947) with boundaries at $(0, h)$.

The risk adjusted CUSUM scheme repeatedly tests the hypotheses given by

$$H_0 : \pi = \pi_0 \quad \text{such that } \frac{\pi_0}{1-\pi_0}/\frac{p_t}{1-p_t} = R_0 \quad \text{and}$$

$$H_A : \pi = \pi_1 \quad \text{such that } \frac{\pi_1}{1-\pi_1}/\frac{p_t}{1-p_t} = R_A$$

(5.11)

where $\pi$ is the probability of the outcome of interest for patient $t$, for $t = 1, 2, \ldots$, $p_t$ is the expected probability of the outcome predicted by an appropriate risk model, and $R_0$ and $R_A$ are constant odds ratios. Thus, the decision interval form of upward risk adjusted CUSUM schemes is given by

$$C_t = \max \left\{ 0, C_{t-1} + \gamma \left( y_t + \frac{\Delta_t}{\gamma} \right) \right\}, \quad R_0 < R_A$$

(5.12)

and for the downward scheme by

$$C_t = \min \left\{ 0, C_{t-1} - \gamma \left( y_t + \frac{\Delta_t}{\gamma} \right) \right\}, \quad R_0 > R_A,$$

(5.13)

where $y_t \in \{0, 1\}$ and $\gamma$ and $\Delta$ are given by

$$\gamma = \log \left\{ \frac{R_A}{R_0} \right\} \quad \text{and} \quad \Delta_t = \log \left\{ \frac{1 - (1 - R_0)p_t}{1 - (1 - R_A)p_t} \right\}.$$  

The difference between the risk adjusted scheme and the fixed rate scheme (Equation 5.9) is that, because the tuning factor $\Delta_t$ of the risk adjusted CUSUM is a
function of the expected outcome rate $p_t$, the weight $W_t$ as defined in Equation (5.4) varies as $p_t$ varies.

The plots in Figure 5.5 illustrate the relationship between $p_t$ and $W_t$ for the events $y_t = 0$ and $y_t = 1$, where $R_0 = 1$ and $R_A = 2$ or $\frac{1}{2}$. They show that, as $p_t$ increases the magnitude of the negative weight, $W_t$, for the event $y_t = 0$, increases, but the positive $W_t$, for the event $y_t = 1$, decreases. In effect, a risk adjusted CUSUM monitoring scheme imposes a higher penalty if a lower risk patient experiences an adverse event, but gives a higher reward for a satisfactory outcome for a higher risk patient.

### 5.3.3 Performance of the Risk Adjusted CUSUM as Risk Profile of Population Varies

The risk adjusted CUSUM schemes studied in this section were tuned to detect a shift from $R_0 = 1$ to $R_A = 2$ with decision thresholds $h$ of 4.5, 3.5, and 2.5.

The R, Version 2.0.1, statistical package was used to simulate $\text{ARL}_0$s and $\text{ARL}_1$s. For the cardiac surgery populations, the associated risk score $X_t$ for each patient $t$ was drawn at random from one of the distributions in Figures 5.3 and 5.4 and the expected mortality rate $p_t$ was computed using the inverse transform of the logit($p_t$) found using Equation (5.10). For the ICU population the expected probability of mortality $p_t$ for each patient $t$ was drawn at random and with replacement from the ICU data set. When simulating in-control run lengths, an event 0 or 1 for each patient $t$ was generated with probability $1 - p_t$ or $p_t$, respectively. For out-of-control run lengths, an event 0 or 1 was simulated with probability $(1 - p_t)/(1 + p_t)$ or $2p_t/(1 + p_t)$, respectively. The simulated outcomes and expected mortality rates were used to compute the weights of the risk adjusted CUSUM scheme given by Equation (5.12). The CUSUM value $C_t$ was updated for each $t$ until $C_T \geq h$. Simulation then ceased and the run length $T$ was recorded.

For each study, this simulation was repeated 100,000 times giving a realization of run lengths from which the approximate average run length was computed. This simulation procedure assumes that the true $p_t$ is given by Equation (5.10) and the risk score $X_t$. The effect of the error in the estimation of $p_t$ on the CUSUM
Figure 5.5: Examples showing how the weights $W_t$ of a risk adjusted CUSUM scheme vary with the expected probabilities of mortality $p_t$ predicted by the risk score. Plot (a), where the scaling factor $\gamma = 2$, is for an upward CUSUM and Plot (b), where $\gamma = \frac{1}{2}$, is for a downward CUSUM.
Table 5.1: Results of simulation studies to approximate ARL$_0$s of a risk adjusted CUSUM scheme with $R_A = 2$ used to monitor the cardiac surgical and ICU patient populations described in Section 5.3.1. Note that the ARL$_0$s were approximated from 100,000 simulated runs and their value increases as the average probability of mortality $\bar{p}$ of the patient population decreases.

<table>
<thead>
<tr>
<th>$p^\dagger$</th>
<th>$h^* = 4.5$</th>
<th>$h^* = 3.5$</th>
<th>$h^* = 2.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARL$_0$</td>
<td>C.I.$^\ddagger$</td>
<td>ARL$_0$</td>
</tr>
<tr>
<td>0.007</td>
<td>51595 (51280, 51910)</td>
<td></td>
<td>17750 (17642, 17857)</td>
</tr>
<tr>
<td>0.009</td>
<td>41162 (40910, 41414)</td>
<td></td>
<td>14205 (14119, 14290)</td>
</tr>
<tr>
<td>0.017</td>
<td>23560 (23416, 23704)</td>
<td></td>
<td>8112 (8063, 8161)</td>
</tr>
<tr>
<td>0.020</td>
<td>20655 (20529, 20780)</td>
<td></td>
<td>7069 (7026, 7112)</td>
</tr>
<tr>
<td>0.024</td>
<td>17804 (17695, 17913)</td>
<td></td>
<td>6143 (6106, 6180)</td>
</tr>
<tr>
<td>0.030</td>
<td>12605 (12528, 12682)</td>
<td></td>
<td>4323 (4297, 4349)</td>
</tr>
<tr>
<td>0.034</td>
<td>11016 (10949, 11083)</td>
<td></td>
<td>3793 (3771, 3816)</td>
</tr>
<tr>
<td>0.046</td>
<td>8783 (8730, 8835)</td>
<td></td>
<td>3025 (3007, 3044)</td>
</tr>
<tr>
<td>0.051</td>
<td>8003 (7954, 8052)</td>
<td></td>
<td>2736 (2719, 2753)</td>
</tr>
<tr>
<td>0.058</td>
<td>7071 (7027, 7114)</td>
<td></td>
<td>2429 (2415, 2444)</td>
</tr>
<tr>
<td>0.160</td>
<td>5379 (5346, 5412)</td>
<td></td>
<td>1869 (1857, 1880)</td>
</tr>
</tbody>
</table>

$^*$Decision Interval  $^\dagger$Average probability of mortality  $^\ddagger$95% Confidence Interval

Performance will be examined in Chapter 6.

Table 5.1 gives a summary of the approximate ARL$_0$s and associated 95% confidence intervals when the risk adjusted CUSUM scheme is used to monitor the eleven patient populations described in Section 5.3.1. For each value $h$ of the decision threshold, there is a monotone increase in ARL$_0$ as the average expected mortality rate of the patient population decreases. The plots in Figure 5.6 show that this increase is more pronounced when expected mortality rates are under 5%. The summaries in Table 5.2 and plots in Figure 5.7 show that, as the average risk of mortality of the population changes, there are similar changes for ARL$_1$s.
Figure 5.6: Plots illustrating the increase in the $\text{ARL}_0$ of risk adjusted CUSUM schemes as the average expected probability of mortality $\bar{p}$ of the population being monitored decreases. Note the marked rate of increase in $\text{ARL}_0$ as $\bar{p}$ decreases, for $\bar{p} < 0.05$. 
Table 5.2: Results of simulation studies to approximate ARLs of a risk adjusted CUSUM scheme with $R_A = 2$ used to monitor the cardiac surgical and ICU patient populations described in Section 5.3.1. Note that the ARLs were approximated from 100,000 simulated runs and their value increases as the average probability of mortality $\bar{p}$ of the patient population decreases.

<table>
<thead>
<tr>
<th>$p$ $\dagger$</th>
<th>$h^* = 4.5$</th>
<th></th>
<th>$h^* = 3.5$</th>
<th></th>
<th>$h^* = 2.5$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARL$_1$</td>
<td>C.I. $\ddagger$</td>
<td>ARL$_1$</td>
<td>C.I. $\ddagger$</td>
<td>ARL$_1$</td>
<td>C.I. $\ddagger$</td>
</tr>
<tr>
<td>0.007</td>
<td>1432.1 (1426.3, 1437.8)</td>
<td></td>
<td>1063.7 (1059.1, 1068.3)</td>
<td></td>
<td>693.6 (690.3, 696.8)</td>
<td></td>
</tr>
<tr>
<td>0.009</td>
<td>1156.8 (1152.2, 1161.5)</td>
<td></td>
<td>851.9 (848.2, 855.6)</td>
<td></td>
<td>559.6 (556.9, 562.2)</td>
<td></td>
</tr>
<tr>
<td>0.017</td>
<td>664.5 (661.9, 667.2)</td>
<td></td>
<td>491.5 (489.4, 493.6)</td>
<td></td>
<td>322.0 (320.5, 323.5)</td>
<td></td>
</tr>
<tr>
<td>0.020</td>
<td>584.5 (582.2, 586.8)</td>
<td></td>
<td>464.2 (462.4, 466.0)</td>
<td></td>
<td>282.4 (281.1, 283.7)</td>
<td></td>
</tr>
<tr>
<td>0.024</td>
<td>503.7 (501.7, 505.7)</td>
<td></td>
<td>373.3 (371.7, 374.9)</td>
<td></td>
<td>243.7 (242.5, 244.8)</td>
<td></td>
</tr>
<tr>
<td>0.030</td>
<td>352.1 (350.7, 353.5)</td>
<td></td>
<td>260.5 (259.4, 261.6)</td>
<td></td>
<td>169.9 (169.1, 170.7)</td>
<td></td>
</tr>
<tr>
<td>0.034</td>
<td>309.6 (308.3, 310.8)</td>
<td></td>
<td>228.5 (227.5, 229.5)</td>
<td></td>
<td>150.3 (149.6, 151.0)</td>
<td></td>
</tr>
<tr>
<td>0.046</td>
<td>250.1 (249.1, 251.1)</td>
<td></td>
<td>183.2 (182.5, 184.0)</td>
<td></td>
<td>120.8 (120.3, 121.4)</td>
<td></td>
</tr>
<tr>
<td>0.051</td>
<td>240.7 (239.8, 241.7)</td>
<td></td>
<td>178.4 (177.7, 179.2)</td>
<td></td>
<td>116.2 (115.6, 116.7)</td>
<td></td>
</tr>
<tr>
<td>0.058</td>
<td>201.2 (200.4, 201.9)</td>
<td></td>
<td>148.8 (148.1, 149.4)</td>
<td></td>
<td>97.6 (97.2, 98.1)</td>
<td></td>
</tr>
<tr>
<td>0.160</td>
<td>163.2 (162.6, 163.8)</td>
<td></td>
<td>121.1 (120.6, 121.6)</td>
<td></td>
<td>79.9 (79.6, 80.3)</td>
<td></td>
</tr>
</tbody>
</table>

* Decision Interval  $\dagger$ Average probability of mortality  $\ddagger$ 95% Confidence Interval
CHAPTER 5: RA-CUSUM—PREDICTION KNOWN

Figure 5.7: Plots illustrating the increase in the $ARL_1$s of risk adjusted CUSUM schemes as the average expected probability of mortality $\bar{p}$ of the population being monitored decreases. Note the mark rate of increase of $ARL_1$s as $\bar{p}$ decreases, for $\bar{p} < 0.05$
Suppose cardiac surgeons perform 200 operations per year. This is the maximum annual average of the number of operations undertaken by any surgeon in the study by Bridgewater et al. (2003). If the decision threshold of an upward risk adjusted CUSUM is set to 4.5, as Steiner et al. (2000) do in their example, the expected time for detection of a doubling of the odds ratio of the expected mortality is 1.25, 3.32, and 7.16 years when monitoring cardiac surgical populations with average expected mortality rates of 0.046, 0.017, and 0.009, respectively. The risk adjusted CUSUM scheme in these studies is tuned to detect a shift of \( \frac{R_A p}{1-p} \approx R_A p \).

In absolute terms it is tuned to detect shifts of approximately 0.042, 0.016, or 0.007 in the average observed mortality rates when the average expected mortality rates are 0.046, 0.017, or 0.007. When using this scheme to monitor cardiac surgical mortality for which the mean expected death rate is likely to be below 0.02, we observe that it takes an unacceptable length of time to detect an increase of less than four post-operative deaths per year. It is suggested that, if the CUSUM signalled such a small increase in the mortality rate, the investigators would be unable to establish that the surgeon’s processes and procedures were truly out-of-control or that the change was simply due to chance. On the other hand, the risk adjusted CUSUM is tuned to detect a shift in the observed mortality rate of the ICU population of approximately 0.116. Given that the ICU at the Princess Alexandra Hospital admitted an average of approximately 1,130 patients per year between 1995 and 1997, the expected time taken for the risk adjusted CUSUM scheme, with \( h = 4.5 \), to detect shift of that magnitude in the observed mortality rate is 53 days. If the ICU was only treating 200 patients per year the expected time taken to signal that the process is out-of-control is 0.82 years.

One way of controlling the increasing \( \text{ARL}_0 \)s of risk adjusted CUSUM is to decrease the decision interval \( h \) as the mean expected probability of mortality \( \bar{p} \) of the patient population decreases. However, adjusting \( h \) to suit \( \bar{p} \) seems contrary to the purpose of the risk adjusted CUSUM which is, as stated by Steiner et al. (2000), a logical way to accumulate evidence over many patients, while adjusting for a changing mix of patient characteristics that significantly affect the risk. Furthermore, Steiner et al. conclude that this is particularly important when monitoring outcomes of surgery at referral centres where referral patterns may
change over time. These stated objectives and the deteriorating performance of the risk adjusted CUSUM when expected outcome rates are small motivated the detailed analysis of the properties of the risk adjusted CUSUM scheme presented in Section 5.4.

5.4 The Properties of the Risk Adjusted CUSUM Schemes

Two hypotheses were considered to explain the results of the simulation studies in Section 5.3.3. For the first which is presented in Section 5.4.1, it was proposed that the changes in the mean and standard deviation of the CUSUM weight as the expected probability of an adverse outcome increases are related to changes in the ARLs. Elementary calculus is used to analyse the relationships between the mean $E(W)$ and the variance $\text{Var}(W)$ of the weight and the expected probability of an adverse outcome $p$ and the tuning factors $R_0$ and $R_A$ of the Steiner CUSUM scheme.

A second, more credible hypothesis is presented in Section 5.4.2. It is that the ARLs of any CUSUM scheme is related to the size of the shift it is tuned to detect. It is suggested that the ARLs of the Steiner CUSUM scheme increases as the mean expected probability of the outcome $\bar{p}$ decreases because the mean of the absolute value of the step size $|\pi_1 - \pi_0|$ the CUSUM scheme is tuned to detect decreases as $\bar{p}$ decreases.
5.4.1 Properties of the Mean and Variance of risk adjusted CUSUM Weights

The Expectation $E(W)$

The mean $E(W)$ of the risk adjusted CUSUM weight is given by

$$E(W) = \sum_{y=0}^{1} \gamma \left( y + \frac{\Delta}{\gamma} \right) p^y (1-p)^{1-y}$$

$$= \gamma \left( \frac{\Delta}{\gamma} + p \right)$$

$$= \log \left\{ \frac{1 - (1 - R_0)p}{1 - (1 - R_A)p} \right\} + p \log \left\{ \frac{R_A}{R_0} \right\}, \quad (5.14)$$

Note that, as $p \to 0$ and as $p \to 1$, $E(W) \to 0$.

Derive Local Minimum of $E(W)$

The first derivative of $E(W)$ with respect to $p$, is given by

$$\frac{d\{E(W)\}}{dp} = \frac{R_0 - R_A}{1 - (2 - R_0 - R_A)p + (1 - R_0)(1 - R_A)p^2} + \log \left\{ \frac{R_A}{R_0} \right\},$$

Consider the case $R_0 = 1$.

There is a stationary point at

$$0 < p = \frac{1}{1 - R_A} + \frac{1}{\log(R_A)} < 1, \quad R_A > 0, R_A \neq 1.$$

The second derivative of $E(W)$ with respect to $p$ is given by

$$\frac{d^2\{E(W)\}}{dp^2} = \frac{(1 - R_A)^2}{\left\{1 - (1 - R_A)p\right\}^2} > 0, \quad p \in (0, 1).$$

Therefore, $E(W)$ has a local minimum at

$$(p, E(W)_{\text{min}}) = \left( \frac{1}{1 - R_A} + \frac{1}{\log(R_A)}, \frac{\log(R_A)}{1 - R_A} + 1 - \log \left\{ \frac{R_A - 1}{\log(R_A)} \right\} \right),$$

and $E(W) < 0$ for all $p \in (0, 1)$.

$E(W)_{\text{min}}$ as a Function of $R_A$

For the trivial case $R_A = 1$, $E(W) = 0$ for all $p \in (0, 1)$.

As $R_A$ decreases from 1 to 0, $p_{\text{min}}$ increases from $1/2$ to 1 and $E(W)_{\text{min}}$ decreases
from 0 to $-\infty$.

As $R_A$ increases from 1 to $\infty$, $p_{\text{min}}$ decreases from $1/2$ to 0 and $E(W)_{\text{min}}$ decreases from 0 to $-\infty$.

**The Variance** $\text{Var}(W)$

The variance of the risk adjusted CUSUM weight $\text{Var}(W_t)$ is given by

$$
\text{Var}(W) = E(W^2) - [E(W)]^2 \\
= \sum_{y=0}^{1} \left[ \left\{ \gamma(y + \frac{\Delta}{\gamma}) \right\}^2 p^y(1-p)^{1-y} \right] - \left\{ \gamma \left( \frac{\Delta}{\gamma} + p \right) \right\}^2 \\
= \gamma^2 p(1-p) \\
= \log^2 \left\{ \frac{R_A}{R_0} \right\} p(1-p).
$$

(5.15)

By inspection, $\text{Var}(W) \rightarrow 0$ at the boundaries of the domain $p \in (0,1)$ and has a local maximum at $p = 1/2$. Thus, the variance of the risk adjusted CUSUM weight $\text{Var}(W)$ varies parabolically as $p$ varies with its maximum value $\gamma^2/4$ determined by the scaling factor of the risk adjusted CUSUM scheme.

**Results**

The plots of the mean $E(W)$ and variance $\text{Var}(W)$ of the weight of the Steiner CUSUMs with tuning factors $R_A = \frac{2}{3}, 2, 3$ in Figure 5.8 (a) and (b), respectively, and for $R_A = \frac{2}{3}, \frac{1}{2}, \frac{1}{3}$ in Figure 5.8 (c) and (d) provide examples of the behaviour of $E(W)$ and $\text{Var}(W)$ as the expected probability $p$ varies, for $0 < p < 1$. In Chart (a), the minimum value of $E(W)$ decreases as $R_A$ increases and the value of $p$ at which the minimum occurs drifting towards 0 as $R_A (>1)$ increases and, in Chart (b), the maximum value of $\text{Var}(W)$ increases as $R_A (>1)$ increases. The plots in Charts (c) and (d) show similar behaviour of $E(W)$ and $\text{Var}(W)$ when $R_A (<1)$ decreases except that the value of $p$, at which minimum $E(W)$ occurs, drifts towards 1.

Table 5.3 shows that the standard deviation $\text{SD}(W)$ is greater than its mean when $R_A$ takes an order of values expected for risk adjusted CUSUM schemes used to monitor adverse medical outcomes. Results from the study described in Section 5.3 show that the $\text{ARL}_0$ and $\text{ARL}_1$ of risk adjusted CUSUMs increase as
Figure 5.8: Examples to illustrate the changes in the mean $E(W)$ and variance $\text{Var}(W)$ of the risk adjusted CUSUM weights as the expected probability of mortality $p$ varies.

(a) For $R_A > 1$, minimum of mean is skewed to left and decreases as $R_A$ increases.

(b) For $R_A > 1$, maximum variance (at $p = \frac{1}{2}$) increases as $R_A$ increases.

(c) For $R_A < 1$, minimum of mean is skewed to right and decreases as $R_A$ decreases.

(d) For $R_A < 1$, variance increases as $R_A$ decreases.

Note that the plot of $\text{Var}(W)$ for each $R_A$ in Chart (b) is identical the plot of $\text{Var}(W)$ for each respective $1/R_A$ in Chart (d).
Table 5.3: Minimum means $E(W)$ and maximum standard deviations $SD(W)$ of the risk adjusted CUSUM scheme as the odds ratio, $R_A$, to which it is tuned, diverges from one. Note that, for $R_A$ near one, $|E(W)| < SD(W)$, but, as $R_A$ diverges from one, $|E(W)|$ increases more rapidly than $SD(W)$. At $R \approx 200\left(\frac{1}{200}\right)$, $|E(W)| > SD(W)$.

<table>
<thead>
<tr>
<th>$R_A$</th>
<th>$p^*$</th>
<th>$E(W)$</th>
<th>SD(W)</th>
<th>$R_A$</th>
<th>$p^*$</th>
<th>$E(W)$</th>
<th>SD(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{3}{2}$</td>
<td>0.466</td>
<td>-0.021</td>
<td>0.203</td>
<td>$\frac{2}{3}$</td>
<td>0.534</td>
<td>-0.021</td>
<td>0.203</td>
</tr>
<tr>
<td>2</td>
<td>0.443</td>
<td>-0.060</td>
<td>0.347</td>
<td>$\frac{1}{2}$</td>
<td>0.557</td>
<td>-0.060</td>
<td>0.347</td>
</tr>
<tr>
<td>3</td>
<td>0.410</td>
<td>-0.148</td>
<td>0.549</td>
<td>$\frac{1}{3}$</td>
<td>0.590</td>
<td>-0.148</td>
<td>0.549</td>
</tr>
<tr>
<td>10</td>
<td>0.323</td>
<td>-0.619</td>
<td>1.151</td>
<td>$\frac{1}{10}$</td>
<td>0.677</td>
<td>-0.619</td>
<td>1.151</td>
</tr>
<tr>
<td>100</td>
<td>0.207</td>
<td>-2.114</td>
<td>2.303</td>
<td>$\frac{1}{100}$</td>
<td>0.793</td>
<td>-2.114</td>
<td>2.303</td>
</tr>
<tr>
<td>197</td>
<td>0.184</td>
<td>-2.641</td>
<td>2.642</td>
<td>$\frac{1}{197}$</td>
<td>0.816</td>
<td>-2.641</td>
<td>2.642</td>
</tr>
<tr>
<td>198</td>
<td>0.184</td>
<td>-2.645</td>
<td>2.644</td>
<td>$\frac{1}{198}$</td>
<td>0.816</td>
<td>-2.645</td>
<td>2.644</td>
</tr>
<tr>
<td>1000</td>
<td>0.144</td>
<td>-3.981</td>
<td>3.454</td>
<td>$\frac{1}{1000}$</td>
<td>0.856</td>
<td>-3.981</td>
<td>3.454</td>
</tr>
</tbody>
</table>

$p^*$ at which minimum $E(W)$ occurs
the risk profile of the patient population decreases. It was hypothesized that this behaviour is due to the increase in SD(W) as \( p \) increases. There is increasing negative bias from \( E(W) \) as \( R_A \) diverges from 1 but it was assumed that the effect of the negative bias is outweighed by the larger effect of the standard deviation. However, Table 5.3 shows that, as \( R_A \) diverges from 1, the magnitude of the minimum mean of the risk adjusted CUSUM weight increases more rapidly than the maximum value of the standard deviation and surpasses it when \( R_A \approx 200 \).

### 5.4.2 Relationship between the Absolute Effect Size and Expected Probability of Mortality

The null and alternative hypotheses in Equation (5.11) are expressed in relative terms as constants. Define the effect size \( \Upsilon \) as \( \Upsilon = |\pi_1 - \pi_0| \), where \( \pi = \pi_1 \) is probability of an adverse outcome for patient \( t \) if the process is out-of-control and \( \pi = \pi_0 \) is the probability of an adverse outcome if the process is in-control. For \( R_0 = 1 \), \( \pi_0 = p \), where \( p \) is the expected probability of an adverse outcome for patient \( t \). The value of \( \Upsilon \) is given in terms of \( R_A \) and \( p \) by

\[
\Upsilon(p) = \begin{cases} 
\frac{(R_A - 1)p(1 - p)}{1 + (R_A - 1)p}, & R_A > 1 \\
\frac{(R_A - 1)p(1 - p)}{1 + (R_A - 1)p}, & 0 < R_A < 1
\end{cases}
\]  

(5.16)

**Derive Local Minimum of \( \Upsilon \)**

Equation (5.16) has first derivative with respect to \( p \) of

\[
\frac{d\Upsilon}{dp} = \begin{cases} 
\frac{\{R_A - 1\}(1 - 2p - (R_A - 1)p^2)}{\{1 + (R_A - 1)p\}^2}, & R_A > 1 \\
\frac{\{R_A - 1\}(1 - 2p - (R_A - 1)p^2)}{\{1 + (R_A - 1)p\}^2}, & 0 < R_A < 1
\end{cases}
\]

with stationary point on the domain \( p \in (0, 1) \) at

\[
0 < p = \frac{1}{1 + \sqrt{R_A}} < 1, \quad R_A > 0, R_A \neq 1.
\]  

(5.17)
The second derivative of $\Upsilon(p)$ (Equation 5.16) with respect to $p$ is

$$
\frac{d^2 \Upsilon}{dp^2} = \begin{cases} 
-\frac{2R_A(R_A - 1)(1 - p + R_A p)}{[1 + (R_A - 1)p]^4}, & R_A > 1 \\
\end{cases}
$$

Clearly $\Upsilon'' < 0$ when $p \in (0, 1)$, therefore the stationary points defined in Equation (5.17) are local maxima.

**Shape of $\Upsilon$**

From Equation (5.16), $\Upsilon \to 0$ as $p \to 0$ or $p \to 1$

$\Upsilon$ has a maximum value $\Upsilon_{\text{max}} = \left| \left( \sqrt{R_A} - 1 \right) / (\sqrt{R_A} + 1) \right|$.

$\Upsilon > 0$ for all $p \in (0, 1)$.

**$\Upsilon$ as a Function of $R_A$**

$\Upsilon \to 1$, the maximum possible effect size, as $R_A \to \infty$ or $R_A \to 0$.

As $R_A > 1$ increases the probability $p_{\text{max}}$, at which $\Upsilon_{\text{max}}$ occurs, drifts from $1/2$ to 0.

As $R_A < 1$ decreases $p_{\text{max}}$ drifts from $1/2$ to 1.

**Discussion**

The analysis presented in this section led to the formulation of the hypothesis that ARLs of the risk adjusted CUSUM is related to the effect size $|\pi_1 - \pi_0|$ to which it is tuned to detect. Because the values of $|\pi_1 - \pi_0|$ for $p < 0.5$ are less for downward risk adjusted CUSUM schemes, where $R_A < 1$, than upward schemes, where $R_A > 1$, this hypothesis provides an explanation for both the upward and downward charts in Steiner et al. (2000, Example) having ARL0s of around 9,600, but decision thresholds set at 4.5 and −4.0, respectively. The plots in Figure 5.9 shows that the effect size of an upward risk adjusted CUSUM scheme with tuning factor $R_A = 2$ is greater than the effect size of a downward CUSUM scheme with tuning factor $R_A = \frac{1}{2}$, for $p < \frac{1}{2}$. Consequently, if the decision thresholds were equal, one would expect the ARL0 of an upward scheme to be less than that of a downward scheme because risk adjusted CUSUMs are used to monitor medical
Figure 5.9: Plots showing the change in the absolute effect sizes $\Upsilon(p) = |\pi_1 - \pi_0|$ of an upward risk adjusted CUSUM scheme tuned to detect a doubling in the odds ratio of observed to expected outcomes ($R_A = 2$) and a downward scheme tuned to detect halving of the odds ratio ($R_A = \frac{1}{2}$) as the expected probability of mortality predicted by the risk score, $p$, varies.

outcomes where the expected probability of an adverse event is usually $p < \frac{1}{2}$.

Support for this hypothesis is provided by Jennison and Turnbull (2006). They note that the average sample number of sequential tests using SPRT decreases as the treatment effect size increases. Because the risk adjusted CUSUM scheme is a sequence of SPRT tests, the behaviour of the effect size $\Upsilon(p)$ offers a credible explanation for the variations of $\text{ARL}_0$s and $\text{ARL}_1$s seen in the results of the simulation study given in Section 5.3. Accordingly, both a modification of the risk adjusted CUSUM scheme to stabilize $\Upsilon(p)$ is proposed and the results of simulation studies to determine its ARLs as $p$ varies are presented in Section 5.5.
5.5 Comparison of the Risk Adjusted and Minimum Effect CUSUM Schemes

The modified risk adjusted CUSUM scheme proposed in this section is tuned to detect a shift from the probability of an adverse event for an in-control process $\pi_{0,t}$ to the probability of an adverse event for an out-of-control process $\pi_{1,t}$ such that the shift $|\pi_{1,t} - \pi_{0,t}| \geq 0.05$ for all patients $t$. The shift $|\pi_{0,t} - \pi_{1,t}|$ will be called the effect size and the modified risk adjusted CUSUM scheme will be called the minimum effect CUSUM scheme.

There are an infinite number of possible designs for a minimum effect CUSUM scheme. The CUSUM design chosen for the simulation studies in this section is given by

$$
\pi_1 = \begin{cases} 
0.05 + 1.5p, & p < 0.1 \\
\frac{8(R_A - 1)}{R_A + 4} p + \frac{1.6 - 0.6R_A}{R_A + 4}, & 0.1 \leq p < 0.2 \\
\frac{R_A p}{1 + (R_A - 1)p}, & p \geq 0.2
\end{cases}
$$

(5.18)

where $\pi_1$ is the out-of-control parameter value, $p$ is the expected outcome rate, and $R_A$ is the out-of-control odds ratio used in the Steiner CUSUM scheme. The value $\pi_0$ of the parameter $\pi$ when the process is in-control is $p$.

When discussing the design of CUSUM schemes used in the industrial context, Hawkins and Olwell (1998, page 32) note that schemes should be tuned to detect a shift that would have a meaningful impact on the process operation but be small enough not to be obvious. The minimum effect CUSUM scheme in Equation (5.18) appears to meet these criteria. The value of 0.05 for the minimum effect size was chosen because it was the value mentioned by a clinical practitioner during an in-house workshop.

The plots in Figure 5.10 compare the effect size $|\pi_1 - \pi_0|$ as $p$ varies, of the proposed CUSUM scheme given in Equation (5.18) with that of the original risk adjusted CUSUM scheme (Steiner et al., 2000). For the original scheme, the effect size falls below 0.05 where $p$ is less than 0.053 or greater than 0.947, approximately, but for the new scheme, it only falls below 0.05 where $p > 0.947$. The proposed
Figure 5.10: Plots showing the change in the absolute effect size $\Upsilon = |\pi_1 - \pi_0|$, to which the minimum effect scheme is tuned, if the effect size is constrained using the scheme given in Equation (5.18) compared with the change in the absolute effect size given by the risk adjusted CUSUM scheme, where $R_A = 2$, as the expected probability of mortality $p$ varies.
scheme remains a risk adjusted CUSUM, because \( \pi_0 \) and \( \pi_1 \) depend on \( p \). If it is used to monitor medical processes, which typically have small adverse outcome rates, it will be tuned to detect clinically meaningful shifts in the observed outcome rates.

Simulation studies to find the approximate ARL\(_0\)s and ARL\(_1\)s of the minimum effect CUSUM scheme given in Equation (5.18) were undertaken to assess its performance for the eleven patient populations used to study the original risk adjusted CUSUM scheme. The simulation procedure used was similar to that outlined in Section 5.3.3 except that the events 0 or 1 for the process out-of-control were simulated using the parameter values \( \pi_1(p_t) \) given by the rule in Equation (5.18).

The approximate ARL\(_0\)s and 95% confidence intervals for each patient population are given in Tables 5.4. For the cardiac surgical populations, the ARL\(_0\)s for \( h = 4.5 \) decrease from 4,545 to 3,725 as the average expected mortality rate increases from 0.009 to 0.058. The relatively small changes in the ARL\(_0\)s of the minimum effect CUSUM scheme as \( \bar{p} \) varies are less than for the original risk adjusted CUSUM scheme. Table 5.1 shows that, when \( h = 4.5 \), its ARL\(_0\)s decrease from 41,162 to 7,071 as \( \bar{p} \) increases from 0.009 to 0.058. Comparison of the ARL\(_0\)s, with the decision interval \( h \) set to 3.5 or 2.5, in Table 5.4 with those in Table 5.1 confirm that the ARL\(_0\)s of minimum effect risk adjusted CUSUM scheme appear more stable than the ARL\(_0\)s of the original risk adjusted scheme. It appears that the design of the minimum effect CUSUM scheme has achieved its goal of stabilising the ARLs of a risk adjusted CUSUM scheme used to monitor a patients where the risk profile of the patient population changes.

Table 5.5 gives the ARL\(_1\)s and associated confidence intervals for the minimum effect CUSUM scheme where the decision thresholds are \( h = 4.5, 3.5 \) and 2.5. For \( h = 4.5 \) the ARL\(_0\)s when monitoring the cardiac surgical populations increase slowly from 51.4 to 88.6 as the expected risk of the patient population increases. When the ARL\(_1\)s in Tables 5.2 and 5.5 are compared, the ARL\(_1\)s for minimum effect CUSUM schemes also appear smaller and more stable, but this is expected as the minimum effect ARL\(_0\)s were also observed to be smaller than those for the original risk adjusted CUSUM schemes.

The plots labelled ‘Min Effect CUSUM’ and ‘RA-CUSUM’ in Figure 5.11 allow
Table 5.4: Results of simulation studies to approximate ARL\(_0\)s of a risk adjusted CUSUM scheme with a minimum effect size of 5\% used to monitor the cardiac surgical and ICU patient populations described in Section 5.3.1. Note that the ARL\(_0\)s were approximated from 100,000 runs and they are more stable than those given in Table 5.1.

<table>
<thead>
<tr>
<th>(p)†</th>
<th>(h^* = 4.5)</th>
<th>C.I.‡</th>
<th>(h^* = 3.5)</th>
<th>C.I.‡</th>
<th>(h^* = 2.5)</th>
<th>C.I.‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.007</td>
<td>4485 (4457, 4512)</td>
<td>1767 (1757, 1778)</td>
<td>455 (452, 458)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.009</td>
<td>4545 (4517, 4573)</td>
<td>1708 (1698, 1719)</td>
<td>472 (469, 475)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.017</td>
<td>4512 (4484, 4540)</td>
<td>1569 (1560, 1579)</td>
<td>524 (520, 527)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.020</td>
<td>4450 (4423, 4477)</td>
<td>1563 (1554, 1573)</td>
<td>514 (511, 517)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.024</td>
<td>4397 (4370, 4425)</td>
<td>1551 (1542, 1561)</td>
<td>511 (507, 514)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.030</td>
<td>4058 (4033, 4083)</td>
<td>1412 (1404, 1421)</td>
<td>473 (471, 476)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.034</td>
<td>3986 (3961, 4010)</td>
<td>1390 (1382, 1399)</td>
<td>452 (449, 454)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.046</td>
<td>3989 (3965, 4013)</td>
<td>1343 (1335, 1352)</td>
<td>432 (430, 435)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.051</td>
<td>3804 (3780, 3827)</td>
<td>1315 (1307, 1323)</td>
<td>422 (419, 424)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.058</td>
<td>3725 (3702, 3748)</td>
<td>1288 (1281, 1296)</td>
<td>408 (406, 411)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.160</td>
<td>3750 (3727, 3773)</td>
<td>1319 (1311, 1327)</td>
<td>429 (426, 431)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Decision Interval † Average probability of mortality ‡ 95\% Confidence Interval
TABLE 5.5: Results of simulation studies to approximate ARL\(_1\)s of a risk adjusted CUSUM scheme with a minimum effect size of 5% used to monitor the cardiac surgical and ICU patient populations described in Section 5.3.1. Note that the ARL\(_1\)s were approximated from 100,000 runs and they are more stable than those given in Table 5.2.

<table>
<thead>
<tr>
<th>(p)</th>
<th>(h^* = 4.5)</th>
<th>C.I.(^\dagger)</th>
<th>(h^* = 3.5)</th>
<th>C.I.(^\dagger)</th>
<th>(h^* = 2.5)</th>
<th>C.I.(^\dagger)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.007</td>
<td>51.4 (51.2, 51.7)</td>
<td>40.6 (40.4, 40.8)</td>
<td>27.6 (27.4, 27.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.009</td>
<td>53.7 (53.4, 53.9)</td>
<td>42.0 (41.8, 42.2)</td>
<td>28.8 (28.7, 29.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.017</td>
<td>61.3 (61.0, 61.6)</td>
<td>47.1 (46.9, 47.4)</td>
<td>33.5 (33.3, 33.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.020</td>
<td>62.8 (62.5, 63.0)</td>
<td>48.4 (48.1, 48.6)</td>
<td>34.2 (34.0, 34.4)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.024</td>
<td>65.0 (64.7, 65.3)</td>
<td>49.9 (49.7, 50.2)</td>
<td>35.4 (35.3, 35.6)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.030</td>
<td>84.6 (84.3, 85.0)</td>
<td>63.7 (63.4, 64.0)</td>
<td>43.6 (43.4, 43.8)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.034</td>
<td>85.9 (85.6, 86.3)</td>
<td>64.8 (64.5, 65.1)</td>
<td>43.9 (43.7, 44.1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.046</td>
<td>86.7 (86.4, 87.1)</td>
<td>64.8 (64.5, 65.1)</td>
<td>43.8 (43.5, 44.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.051</td>
<td>87.6 (87.2, 88.0)</td>
<td>65.6 (65.4, 65.9)</td>
<td>44.1 (43.9, 44.3)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.058</td>
<td>88.6 (88.2, 88.9)</td>
<td>66.5 (66.2, 66.8)</td>
<td>44.5 (44.3, 44.7)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.160</td>
<td>76.7 (76.3, 77.0)</td>
<td>58.4 (58.2, 58.7)</td>
<td>40.7 (40.5, 40.9)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\dagger\) Decision Interval  \(\dagger\) Average probability of mortality  \(\dagger\) 95% Confidence Interval
a comparison of the standardized ARL\(_1\)s for the minimum effect and the original schemes when the ARL\(_0\) is fixed at 5,000 for each of the populations being monitored. The method used to make these comparisons was to fit curves of the form

\[
\text{ARL} = ah^2 + bh + c \tag{5.19}
\]

through the three points defined by the decision thresholds \(h = 4.5, 3.5, \) and 2.5 and their associated approximate ARL\(_0\)s or ARL\(_1\)s given in Tables 5.1, 5.4, 5.2, or 5.5. Each decision threshold \(h'\) to give an ARL\(_0\) of 5,000 was estimated using interpolation for the original risk adjusted CUSUM scheme or extrapolation for minimum effect scheme. The standardized ARL\(_1\) was found by substituting the estimate \(h'\) and the appropriate values of parameters \(a, b,\) and \(c\) into Equation (5.19). The plots show that, as the mean expected probability \(p\) of mortality of populations decreases, the ARL\(_1\)s of the original risk adjusted CUSUM scheme, tuned to detect an upward shift in mortality outcomes, increase as \(p\) decreases with the rate of increase accelerating as \(p \to 0\). On the other hand, the ARL\(_1\)s of the minimum effect CUSUM scheme remain relatively stable for all populations studied.

The plot labelled ‘RA & Min Effect’ gives standardized ARL\(_1\)s for the original risk adjusted CUSUM schemes used to monitor the mortality outcomes of the cardiac surgical and ICU patient populations described previously. For these studies, the observed out-of-control mortality rates are assumed to have exceeded the expected mortality rates according to the rule given in Equation (5.18). In this case, quadratic curves have been fitted to the ARL\(_1\)s given in Table 5.6 and the standardized ARL\(_1\)s calculated by substituting the values \(h'\) estimated for the original risk adjusted CUSUM schemes into Equation (5.19). The plot of the standardized ARL\(_1\)s for the RA & Min Effect studies shows that they are comparable to but greater than the ARL\(_1\)s estimated for the Min Effect CUSUM. The two plots demonstrate the robustness of the CUSUM and are consistent with its optimality properties as discussed in Section 5.2.

The plot for RA-CUSUM in Figure 5.12 shows that to standardize the performance of the original risk adjusted CUSUM scheme by setting the ARL\(_0\)s to
5,000 it necessary to set the decision threshold $h'$ above 4 when monitoring the ICU population, which has a mean expected mortality rate of 0.160, and the cardiac surgical population with mean expected mortality rate of 0.058. On the other hand, $h'$ must be set to less than 2.5 for the cardiac surgical population with mean expected mortality rate of 0.007. By comparison, if the minimum effect risk adjusted CUSUM scheme is tuned using the rule given by Equation (5.18), an ARL$_0$ of 5,000 is given by $h'$ greater than 4.5 for all eleven patient populations being monitored.

The advantage of the minimum effect scheme is that it stabilizes the performance of risk adjusted CUSUMs as the risk profile of the population being monitored varies. The disadvantage is that the tuning rule is more complex than that used for the original risk adjusted CUSUM scheme. However, any tuning rule that limits the effect size of a risk adjusted CUSUM scheme to no less than 0.05, or some other reasonable minimum, will stabilize the CUSUM’s performance. For example, in another simulation study, for which the risk adjusted scheme was tuned to detect an increase in the observed mortality rate of 0.05 when $p < 0.1$, an increase of 0.1 when $0.1 < p < 0.2$, and a doubling of the odds of the observed mortality rate when $p > 0.2$, the candidate found that, for this CUSUM scheme, the values of both the ARL$_0$s and the ARL$_1$s as $\bar{p}$ varied were similar to those of the minimum effect CUSUM scheme given Tables 5.4 and 5.5, respectively. The plots in Figure 5.12 suggest the performance characteristics of the original and minimum effect CUSUM schemes are comparable if the patient populations have an average expected outcome rate greater than 0.05. For example, the CUSUMs in Figures 5.13(a) and (b), which were constructed using the original and a minimum effect risk adjusted schemes, monitor the mortality outcomes in the ICU, Princess Alexandra Hospital, for the three years 1995 to 1997. Their performances, where there was a signal from the minimum effect scheme on 26 June 1995 and one from the original risk adjusted scheme three days later, are comparable.
SECTION 5: MINIMUM EFFECT CUSUM

Figure 5.11: ARL$_1$s plotted against average mortality rates for cardiac surgical and ICU patient populations described in Section 5.3.1 with ARL$_0$s for each CUSUM scheme set to 5,000.

- Min Effect CUSUM: Minimum effect scheme tuned using rule given in Equation (5.18);
- RA-CUSUM: Risk adjusted scheme tuned using rule given in Equation (5.12);
- RA & Min Effect: Risk adjusted scheme tuned according to Equation (5.12), but the probabilities of mortality were taken as the out-of-control parameter values of the Min Effect CUSUM.
Table 5.6: Results of simulation studies to approximate the ARLs of a risk adjusted CUSUM scheme, tuned according to the rule given by Equation (5.12) with $R_A = 2$, but deaths were simulated with probability $\pi_1$ given by the out-of-control parameter in Equation (5.18) for the minimum effect CUSUM scheme.

<table>
<thead>
<tr>
<th>$p$</th>
<th>$h^* = 4.5$</th>
<th>$h^* = 3.5$</th>
<th>$h^* = 2.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARL1</td>
<td>C.I.‡</td>
<td>ARL1</td>
</tr>
<tr>
<td>0.007</td>
<td>134.5 (134.1, 134.8)</td>
<td>105.4 (105.1, 105.7)</td>
<td>76.6 (76.3, 76.8)</td>
</tr>
<tr>
<td>0.009</td>
<td>133.4 (133.0, 133.7)</td>
<td>104.0 (103.7, 104.3)</td>
<td>76.0 (75.8, 76.3)</td>
</tr>
<tr>
<td>0.017</td>
<td>128.8 (128.5, 129.2)</td>
<td>100.5 (100.2, 100.8)</td>
<td>72.4 (72.1, 72.6)</td>
</tr>
<tr>
<td>0.020</td>
<td>127.0 (126.6, 127.3)</td>
<td>99.2 (98.9, 99.5)</td>
<td>71.4 (71.2, 71.7)</td>
</tr>
<tr>
<td>0.024</td>
<td>125.8 (125.4, 126.1)</td>
<td>97.7 (97.4, 98.1)</td>
<td>70.3 (70.1, 70.6)</td>
</tr>
<tr>
<td>0.030</td>
<td>121.8 (121.5, 122.2)</td>
<td>94.6 (94.3, 94.9)</td>
<td>67.0 (66.8, 67.3)</td>
</tr>
<tr>
<td>0.034</td>
<td>119.9 (119.6, 120.3)</td>
<td>92.8 (92.5, 93.1)</td>
<td>65.5 (65.2, 65.7)</td>
</tr>
<tr>
<td>0.046</td>
<td>120.0 (119.6, 120.4)</td>
<td>92.7 (92.4, 93.0)</td>
<td>65.2 (65.0, 65.5)</td>
</tr>
<tr>
<td>0.051</td>
<td>112.9 (112.5, 113.2)</td>
<td>86.6 (86.3, 86.9)</td>
<td>60.5 (60.2, 60.7)</td>
</tr>
<tr>
<td>0.058</td>
<td>110.2 (109.9, 110.6)</td>
<td>84.3 (84.0, 84.6)</td>
<td>58.5 (58.3, 58.8)</td>
</tr>
<tr>
<td>0.160</td>
<td>103.6 (103.2, 103.9)</td>
<td>78.7 (78.5, 79.0)</td>
<td>54.1 (53.9, 54.3)</td>
</tr>
</tbody>
</table>

*Decision Interval †Average probability of mortality ‡95% Confidence Interval
**SECTION 5: MINIMUM EFFECT CUSUM**

Figure 5.12: Decision thresholds $h$ plotted against average mortality rates for cardiac surgical and ICU patient populations described in Section 5.3.1 and for $\text{ARL}_0$s are set to 5,000.

- **Min Effect CUSUM**: Minimum effect scheme tuned using rule given in Equation (5.18);
- **RA-CUSUM**: Risk adjusted scheme tuned using rule given in Equation (5.12).
(a) Risk Adjusted CUSUM

(b) Min Effect CUSUM

**Figure 5.13**: Comparison of a risk adjusted and a minimum effect CUSUM scheme monitoring mortality outcomes in the ICU, Princess Alexandra Hospital. Plot (a) is from a risk adjusted CUSUM tuned to detect a doubling of the observed odds ratio. Plot (b) is from a minimum effect CUSUM tuned according to the rule in Equation (5.18).
5.6 Comparison of Industrial and Risk Adjusted CUSUM Schemes

The results reported in this section provide an interesting contrast between the behaviour of the ARL\(_0\)s of the upward binomial CUSUM scheme given by Equation (5.6) and the upward Steiner CUSUM scheme given by Equation (5.9). The motivation for comparing the ARL\(_0\)s of unscaled binomial CUSUM schemes and scaled Steiner CUSUM schemes was the observation that the ARL\(_0\) of the illustrative example of the upward binomial CUSUM scheme in Figure 5.1 decreased as the effect size that it was tuned to detect decreased. In contrast, the ARL\(_0\)s of scaled CUSUM schemes were found to increase as the effect size decreases. Consequently, an exploratory study was undertaken to acquire some understanding of the performance of the binomial CUSUM scheme as both its effect size and the proportion of non-conforming items for the in-control process being monitored vary. Its performance is compared with that of a scaled CUSUM scheme.

Each CUSUM scheme was tuned to detect shifts from \(R_0 = 1\) to \(R_A = 2, 3,\) and 4, where \(R_0\) and \(R_A\) are as defined in Equation (5.11), and set to signal when the CUSUM \(C_t \geq 4.5\). They were used to monitor ten populations for which the probabilities of adverse events were a constant taking one of the values given in the column labelled \(p\) in Table 5.14. ARL\(_0\)s, which were used to compare the performances of the two schemes, were approximated from 1,000 simulated run lengths found by using a procedure similar to that outlined in Section 5.3.3.

The results in Table 5.7 show that, as the probability of an adverse outcome decreases, the ARL\(_0\) of both schemes increases, but there is a greater increase for the unscaled scheme. The second finding, illustrated by the plots in Figure 5.14, is that the ARL\(_0\) of the scaled scheme decreases, but that of the unscaled scheme increases, as the effect size increases from \(R_A = 2\) to 4.

The suggested reason for the differences in the performance characteristics of the two CUSUM schemes is the difference in calculating the weights for each scheme. The weights of the risk adjusted CUSUM scheme are computed using the sequential probability ratio test (Wald, 1947) as given in Equations (5.12)
TABLE 5.7: Results of an exploratory simulation study to compare and contrast approximate ARL$_0$s of scaled and unscaled Bernoulli CUSUM schemes, as given by Equations (5.9) and (5.6), respectively. Each ARL$_0$ was estimated from 1,000 simulated run lengths.

<table>
<thead>
<tr>
<th>$p^\dagger$</th>
<th>Odds Ratio: 2</th>
<th>Odds Ratio: 3</th>
<th>Odds Ratio: 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scaled</td>
<td>Unscaled</td>
<td>Scaled</td>
</tr>
<tr>
<td>0.007</td>
<td>53082</td>
<td>10766</td>
<td>18683</td>
</tr>
<tr>
<td>0.009</td>
<td>40138</td>
<td>8916</td>
<td>16399</td>
</tr>
<tr>
<td>0.017</td>
<td>20909</td>
<td>4738</td>
<td>8355</td>
</tr>
<tr>
<td>0.020</td>
<td>17907</td>
<td>4254</td>
<td>7439</td>
</tr>
<tr>
<td>0.024</td>
<td>14912</td>
<td>3624</td>
<td>6261</td>
</tr>
<tr>
<td>0.030</td>
<td>12211</td>
<td>2594</td>
<td>4849</td>
</tr>
<tr>
<td>0.034</td>
<td>11283</td>
<td>2396</td>
<td>4192</td>
</tr>
<tr>
<td>0.046</td>
<td>7999</td>
<td>1916</td>
<td>3446</td>
</tr>
<tr>
<td>0.051</td>
<td>7631</td>
<td>1754</td>
<td>3121</td>
</tr>
<tr>
<td>0.058</td>
<td>6803</td>
<td>1523</td>
<td>2554</td>
</tr>
</tbody>
</table>

†Expected Outcome Rate
#### Figure 5.14: Plots illustrating the contrasting behaviour of the approximate ARL$_0$s of CUSUM schemes scaled weights and schemes with unscaled weights, where the weights, $W_t$, are given by Equations (5.4) and (5.5), respectively. Note

(a) the ARL$_0$s of scaled schemes increase as the tuning odds ratio, $R_A$, increases, but

(b) the ARL$_0$s of unscaled schemes decrease as the tuning odds ratio, $R_A$, increases.
and (5.13). The modification for the binomial CUSUM scheme, where the weights as given in Equations (5.6) and (5.7) are computed without a scaling factor, is not a Wald test. The effect of computing weights of Bernoulli CUSUM schemes without the scaling factor on the CUSUM’s performance is reported here because no previous report of this behaviour was found.

5.7 Conclusion and Recommendation

In the industrial production context, the CUSUM schemes to monitor the number of nonconforming items assume that the probability of a nonconforming item is constant if the process is in statistical control. For CUSUM schemes to monitor the number of adverse outcomes following medical procedures, the assumption that the probability of an adverse outcome is constant is inappropriate because the risk that patients will experience an adverse event is variable. The risk adjusted CUSUM, which is a scheme to monitor the number of adverse outcomes following medical procedures, allows for patient variability. Its weight, $W_t$, depends on the predicted probability, $p$, of an adverse event for patient $t$, where $p$ is found using some risk model. In the studies for this chapter, the risk models chosen were the Parsonnet score or EuroSCORE for cardiac surgical patients and the APACHE III score for ICU patients.

The $\text{ARL}_0$ and $\text{ARL}_1$ measures were used to assess the performance of the risk adjusted CUSUM. The finding is that as the average expected probability $p$ of adverse events decreases, both $\text{ARL}_0$ and $\text{ARL}_1$ increase if the decision interval $(0, h)$ is constant. That is, the number of false alarms and the scheme’s power to detect a truly out-of-control process decrease. Figures 5.6 and 5.7 show that, for a decision threshold, the increases in its $\text{ARL}_0$ and $\text{ARL}_1$ as $\bar{p}$ decreases are greater for $p < 0.05$ than for $p > 0.05$. It was hypothesized that the large values of the $\text{ARL}$ of risk adjusted CUSUM schemes for small average $p$ is because it is set to detect a step from $p$ to $R_Ap$, approximately, and it was found that, by limiting the minimum step size of risk adjusted CUSUM schemes to no less than 0.05, the $\text{ARL}_0$ and $\text{ARL}_1$ of the so-called minimum effect CUSUM scheme were stabilized as the average $p$ of the patient populations varied.
The expected probability of an adverse outcome for many medical procedures, for example, cardiac surgical operations, is less than 0.05. The minimum effect CUSUM scheme was found to be stable if used to monitor the number of mortalities following cardiac surgery. It is recommended for monitoring the outcomes of medical conditions and procedures where the expected probability of an adverse event is less than 0.05.
Chapter 6

Performance of Risk Adjusted CUSUM Schemes: Expected Outcome Probability, $p_t$, Uncertain

6.1 Introduction

In the simulation studies of the ARLs of risk adjusted CUSUM schemes in Chapter 5, the assumption was that, although the risk profile of the patient populations in the studies varied, each expected probability of mortality predicted by the risk scores was equivalent to the patient’s true probability of mortality. However, predicted probabilities of mortality are estimates of the risk of death (Lemeshow et al., 1995). The purpose of this chapter is to investigate the effect of two types of uncertainty in the estimates of the probability of patient mortality has on the performance of risk adjusted CUSUM schemes. In the first investigation, variability in the estimates of the risk of mortality of individual patients and its effect on the ARL$_0$ performance measure of risk adjusted CUSUM schemes are studied. In the second, the effect of variability in estimating the parameters of a model for recalibrating the predictions from a risk score to the patient population being monitored on the ARL$_0$s of risk adjusted and minimum effect CUSUM schemes is investigated.
Concerns about between-patient variability of the probabilities of mortalities calculated using the available risk-adjustment models were raised in the mid-1990s. The magnitude of the patient-level noise in predicted probabilities of mortality was shown in a study of cardiac surgical risk scores and in another of risk scores for patients admitted to an intensive care unit (ICU). When comparing expected probabilities from five cardiac surgical risk scores, Orr et al. (1995) found that there was moderate correlation between predictions of the risk of mortality for individual patients, with the Spearman rank correlation (Riffenburgh, 1999, Pages 419 and 425) ranging from 0.60 to 0.72, with the best correlation between a New York State (Hannan et al., 1994) and the northern New England (O’Connor et al., 1992) models. Comparison of the patients’ probabilities of mortality predicted by these risk scores showed that, occasionally, a patient would appear to be at high risk if the expected probability of mortality was computed using one model, but at much lower risk if it was calculated using another model.

In the other study, Lemeshow et al. (1995) compared individual expected probabilities of mortality from the APACHE II (Knaus et al., 1985) and MPM II (Lemeshow et al., 1993) models for the risk of mortality of ICU patients. Although both models were well calibrated for the patient population in the study and gave similar predictive values for dying and for surviving, they did not necessarily agree in their probability estimates for individual patients, even though both used vital status at hospital discharge as the outcome measure.

Normand et al. (1996) used logistic regression to construct a risk model for short term mortality following admission to hospital for treatment of myocardial infarction. As noted in Chapter 4.4.1, they found that, despite the extraordinary amount of clinical data available, the explanatory power of their predictors was low at 27%. They concluded the unexplained variation (73%) must reflect systematic differences in unmeasured or unobserved patient characteristics, differences in quality of care, inadequacy of the fit of the model, or random error. The inclusion of hospital interventions, which may explain variation due to differences in the standard practices for the delivery of health care, reduced the unexplained variation only slightly, from 73% to 67%. Other possible sources of variability, identified by Lilford et al. (2004), are data quality and data definitions.
The effects of random error in the estimated probabilities of individual patients’ mortality and in the parameter estimates of the associated risk models on the ARL0s of risk adjusted CUSUMs are investigated in this chapter. In Section 6.2, five statistical models to generate uncertainty in the risk of mortality and the method used to incorporate that uncertainty into the algorithms for simulating run length distributions of risk adjusted CUSUM schemes are described. The results from those simulation studies are presented in Section 6.3 and discussed in Section 6.4. The effect that random variation in data used for calibrating risk scores to the patient population of interest has on the parameter estimates of risk models, the estimated probabilities of mortality, and the performance of risk adjusted and minimum effect CUSUM schemes is investigated in Section 6.5. Finally, there are some conclusions in Section 6.6.

### 6.2 Models for Patient Level Uncertainty in Risk Scores

Typically, the predicted probability, \( p_m \), that a patient \( t \) will suffer an adverse event following a medical procedure is derived from some risk score. The risk for patient \( t \) is calculated using a finite set of risk factors, \( X \), as the explanatory variables, and risk score parameters, \( \hat{\beta} \), which are fixed. The predicted probability of the adverse event of all patients with risk factors, \( X \), is \( p_m \). However, every patient undergoing a medical procedure is unique with an unknown “true” probability of an adverse outcome, \( p_T \). In this section, the effect of the uncertainty in the prediction of an individual’s outcome on the ARL0s of risk adjusted and minimum effect CUSUM schemes is investigated using simulation studies. In all of the simulation processes, it is assumed that \( p_T \) for any patient \( t \) is given by

\[
p_T = f(p_m, \varepsilon),
\]

where \( \varepsilon \) is a random error variable drawn from some probability distribution. To ensure fair comparison of the ARL0s approximated from each simulation process, the variance, \( \text{Var}(p_T) \), is assumed to be such that \( \text{Var}[\logit(p_T)] \) is a constant \( \sigma^2 \).
for all \( \logit(p_T) \). Approximate ARLs were simulated with the standard deviation, \( \sigma \), set at 0.45, 0.90, or 1.80.

Some justification for the choice of the low value of 0.45 and the high value of 0.90 is provided by the results from the comparative study by Lemeshow et al. (1995). They compared the expected probabilities of mortality of individuals admitted to an intensive care unit predicted by the APACHE II score with those predicted by the MPM II score. They found both were well calibrated to the observed mortality rate of 0.225 (2,532 of 11,320 ICU patients). The mean probability of mortality of an individual patient predicted by MPM II was 0.228 and the mean prediction of APACHE II was 0.235. However, the standard deviation of the difference of the mean predictions was 0.17.

Suppose that patient \( t \) with “true” probability \( p_T \) has expected probabilities \( p_{m_1} \) and \( p_{m_2} \) predicted by the APACHE II and MPM II scores, respectively. If \( \logit(p_T) \) is given by

\[
\logit(p_T) = \logit(p_{m_1}) + \varepsilon_1 = \logit(p_{m_2}) + \varepsilon_2,
\]

where \( \varepsilon_1 \) and \( \varepsilon_2 \) are unbiased with respect to \( \logit(p_{m_1}) \) and \( \logit(p_{m_2}) \), respectively, such that \( \varepsilon_1, \varepsilon_2 \sim N(0, \sigma^2) \), then

\[
\logit(p_{m_1}) - \logit(p_{m_2}) = \varepsilon_2 - \varepsilon_1 \quad \text{and} \quad \text{Var}(\varepsilon_2 - \varepsilon_1) = 2\sigma^2.
\]

An approximation of \( \text{Var}[\logit(p_T)] \), for \( p_T = 0.235 \) is given by the first order Taylor series approximation (Casella and Berger, 1990, Page 329) as

\[
\text{Var}[\logit(p_T)] \approx \frac{1}{2} \left\{ \frac{d}{dp_T} \left[ \log \left( \frac{p_T}{1 - p_T} \right) \right] \right\}^2 \bigg|_{p_T=0.235} \times 0.17^2
\approx 0.4471. \tag{6.2}
\]

That is, Lemeshow et al. (1995) found the approximate standard deviation of \( \logit(p_T) \) for mortality of the ICU patients in their study was 0.669. When selecting the standard deviations for these simulation studies, it was considered that the interval (0.45, 0.90) represents a plausible range of values for the standard deviation \( \sigma \) of the estimates \( \logit(p_T) \). The third very high value of 1.80 was chosen to
highlight that increasing uncertainty in the risk model increases any bias in the probabilities of adverse outcomes predicted by the risk score.

In the simulation studies, the mortality outcomes of the eleven patient populations described in Chapter 5.3.1 are monitored using the risk adjusted CUSUM scheme described in Chapter 5.3.2 and the minimum effect CUSUM scheme described in Chapter 5.5. The difference between the algorithms used for these studies and those described in Chapter 5.3.3 is that each event 0 or 1 is generated with probability $1 - p_t$ or $p_t$, respectively, where $p_t = p_T$, the “true” probability that the patient will die, instead of $p_t = p_m$, the probability of mortality predicted by the risk score.

Simulation of deaths according the true probabilities of mortality, $p_T$, allows an investigation of the effect, if any, of between patient variability on the ARL$_0$ performance measures of risk adjusted CUSUM schemes. Different models were chosen to generate $p_T$ values to illustrate the effect of any bias in $p_m$ with respect to $p_T$ on the ARL$_0$ approximations. The logistic regression model in Equation 5.9 provides an example of the potential for bias in the estimation of $p_m$. In that model, if the estimated value $X_m$ given by the risk score is normally distributed about the true risk score $X_T$, the values of $p_m$ calculated using Equation 5.9 are biased with respect to $p_T$.

There are five statistical models used to generate the $p_T$ values. In the first model, $p_T$ is drawn from a Beta distribution with mean of $p_m$; in the second, the true odds, $p_T/(1 - p_T)$, are drawn from a Gamma distribution with mean of the risk model odds, $p_m/(1 - p_m)$; in the third, the true odds are drawn from a Gamma distribution with mean chosen so that $p_T$ is unbiased with respect to $p_m$; in the fourth, logit($p_T$) is drawn from a Normal distribution with mean logit($p_m$); and, in the fifth, logit($p_T$) is drawn from a Normal distribution where the mean is chosen so that $p_T$ is unbiased with respect to $p_m$. 
6.2.1 Model 1: Beta Distribution—$p_T$ Unbiased

The Model

The distribution of $p_T$ is given by

$$p_T \sim \text{Beta}(\alpha, \beta),$$

(6.3)

where $E(p_T) = p_m$ and $\text{Var}(p_T) = \sigma^2_{p_T}$.

Model Variance

The value of $\sigma^2_{p_T}$ may be calculated from the risk score, $X$, at $X_{p_m} = \logit(p_m)$ and the standard deviation, $\sigma$, of the risk score using the first order Taylor series approximation

$$\sigma^2_{p_T} \approx \left\{ \frac{d \left( \logit^{-1} X \right)}{dX} \right|_{X=X_{p_m}} \right\}^2 \sigma^2 \approx \frac{e^{2X_{p_m}}}{(1 + e^{X_{p_m}})^4} \sigma^2.$$

(6.4)

Model Parameters

The relationship of the parameters $(\alpha, \beta)$ with the mean and variance of $p_T$ is given by Gelman et al. (1995, Appendix A) as

$$E(p_T) = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}(p_T) = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

(6.5)

where $E(p_T) = p_m$ and $\text{Var}(p_T) = \sigma^2_{p_T}$. Set $k = p_m/(1 - p_m)$. Solving Equation (6.5) for $\alpha$ and $\beta$ gives

$$\beta = \frac{k - \sigma^2_{p_T}(k + 1)^2}{\sigma^2_{p_T}(k + 1)^3} \quad \text{and} \quad \alpha = k \beta.$$

(6.6)

Simulation of Outcomes

For algorithms using the model given in Equation (6.3) to simulate runs of risk adjusted CUSUM schemes, values of $p_T$ are randomly drawn from a $\text{Beta}(\alpha, \beta)$ distribution, where $(\alpha, \beta)$ are given by Equation (6.6), using the \texttt{rbeta} function in the R, Version 2.0.1, statistical package. Outcomes $Y \in \{0, 1\}$, where $Y \sim \text{Bernoulli}(p_T)$, were then simulated using the process described in Section 6.2. The probabilities of such events are unbiased with respect to $p_m$ because $E(Y)$ is given by

$$E(Y) = E[E(Y|p_T)] = E(p_T) = p_m.$$
6.2.2 Model 2: Gamma Distribution—Odds Unbiased

The Model
The distribution of Odds$_{p_T} = p_T / (1 - p_T)$ is given by the multiplicative model

\[
\text{Odds}_{p_T} = k \varepsilon, \quad \varepsilon \sim \Gamma(v, v), \quad (6.8)
\]

where $k = p_m / (1 - p_m)$ and $E(\varepsilon) = 1$.

Model Variance
A first order Taylor series approximation of the variance $\text{Var}(\text{Odds}_{p_T})$ is given by

\[
\text{Var}(\text{Odds}_{p_T}) \approx \left\{ \frac{d (e^X)}{dX} \Bigg|_{X=X_{p_m}} \right\}^2 \sigma^2 
\approx e^{2X_{p_m}} \sigma^2. \quad (6.9)
\]

Thus, the approximate value of $\text{Var}(\varepsilon)$ is given by

\[
\text{Var}(\varepsilon) \approx e^{2X_{p_m}} \sigma^2 \frac{k^2}{v^2}. \quad (6.10)
\]

Model Parameter
Gelman et al. (1995) give the relationship between the mean and variance of a Gamma distribution and its parameters $(\alpha, \beta)$ as

\[
E(\cdot) = \frac{\alpha}{\beta} \quad \text{and} \quad \text{Var}(\cdot) = \frac{\alpha}{\beta^2}. \quad (6.11)
\]

Thus, the value of $v$ in Equation (6.8) is given by

\[
v = \frac{k^2}{\sigma^2 e^{2X_{p_m}}}. \quad (6.12)
\]

Simulation of Outcomes
For algorithms where the “true” probability of mortality is simulated using the model given in Equation (6.8), the values of $\varepsilon$ are drawn from a $\Gamma(v, v)$ distribution, where $v$ is given by Equation (6.12), using the \texttt{rgamma} function in the R statistical package; the values of Odds$_{p_T} = k \varepsilon$ calculated; the values of $p_T$ computed as

\[
p_T = \frac{\text{Odds}_{p_T}}{1 + \text{Odds}_{p_T}}; \quad (6.13)
\]

and the outcomes $Y$ simulated using the process described in Section 6.2. In this model, $E(\text{Odds}_{p_T}) = \text{Odds}_{p_m}$, but $E(p_T) \neq p_m$. That is, the probability of the outcome $Y$ is biased with respect to expected probability, $p_m$, of the event predicted by the risk model.
6.2.3 Model 3: Gamma Distribution—\( p_T \) Unbiased

The Model

The “true” odds, Odds_{\( p_T \)}, such that \( E(p_T) = p_m \), are drawn from a Gamma distribution given by

\[
\text{Odds}_{p_T} = k\varepsilon, \quad \varepsilon \sim \text{Gamma}(\alpha, \beta)
\]

\[
= Z, \quad Z \sim \text{Gamma}\left(\alpha, \frac{\beta}{k}\right), \quad (6.14)
\]

where the parameters (\( \alpha, \beta \)) are chosen such that

\[
E(Z) = m(p_m) \quad \text{and} \quad \text{Var}(Z) = e^{2X_{pm}}\sigma^2. \quad (6.15)
\]

Model Mean

The mean, \( m(p_m) \), is unknown but is chosen such that

\[
E[p_T] = E\left[\frac{Z}{1+Z}\right] = p_m. \quad (6.16)
\]

Set \( f(Z) = Z/(1 + Z) \), then \( f(Z, p_m) \) may be approximated by a second order Taylor series as

\[
f(Z, p_m) \approx f(m, p_m) + \frac{\partial f}{\partial Z}\bigg|_m (Z - m) + \frac{1}{2} \frac{\partial^2 f}{\partial Z^2}\bigg|_m (Z - m)^2
\]

\[
= \frac{m}{1 + m} + \frac{1}{(1 + m)^2} (Z - m) - \frac{1}{(1 + m)^3} (Z - m)^2 \quad (6.17)
\]

For unbiased \( p_T \), it is necessary that

\[
E[f(Z, p_m)] \approx \frac{m}{1 + m} - \frac{1}{(1 + m)^3} e^{2X_{pm}}\sigma^2
\]

\[
= p_m \quad (6.18)
\]

Equation (6.18) may be solved for all \( m = m(p_m) \) by minimizing

\[
\frac{m}{1 + m} - \frac{1}{(1 + m)^3} e^{2X_{pm}}\sigma^2 - p_m \quad (6.19)
\]

using the Matlab function \texttt{fminbnd}.

Model Parameters

The parameters, (\( \alpha, \beta/k \)), for the model in Equation (6.14) are given by

\[
\alpha = \frac{m^2}{e^{2X_{pm}}\sigma^2} \quad \text{and} \quad \frac{\beta}{k} = \frac{m}{e^{2X_{pm}}\sigma^2}. \quad (6.20)
\]
SECTION 2: PATIENT LEVEL UNCERTAINTY

Simulation of Outcomes
The procedure to simulate outcomes $Y$ using the model given in Equation (6.14) is similar to that described in Section 6.2.2. The difference is that values for Odds $p_T$ are drawn such that corresponding values of the “true” probability of mortality, $p_T$, are unbiased with respect to $p_m$.

6.2.4 Model 4: Normal Distribution—Logit Unbiased

The Model
The distribution of logit($p_T$) is given by the additive model

$$\text{logit}(p_T) = \text{logit}(p_m) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \quad (6.21)$$

Simulation of Outcomes
To simulate outcomes $Y$ using the model given in Equation (6.21), values of logit($p_T$) are drawn from an $N(\text{logit}(p_m), \sigma^2)$ distribution using the \texttt{rnorm} function in the R statistical package; the values of $p_T$ calculated as

$$p_T = \frac{e^{\text{logit}(p_T)}}{1 + e^{\text{logit}(p_T)}}, \quad (6.22)$$

and the outcomes $Y$ simulated using the process described in Section 6.2. In this model, draws from the distribution for logit($p_T$) are unbiased with respect to logit($p_m$), but the corresponding values of the “true” probabilities of mortality, $p_T$, are biased with respect to $p_m$.

6.2.5 Model 5: Normal Distribution—$p_T$ Unbiased

The Model
The distribution of logit($p_T$) is given by the additive model

$$\text{logit}(p_T) = \text{logit}(p_m) + \varepsilon, \quad \varepsilon \sim N(m, \sigma^2)$$

$$= Z, \quad Z \sim N(\text{logit}(p_m) + m, \sigma^2), \quad (6.23)$$

where $m = m(p_m)$ is chosen such that $E(p_T) = p_m$.

Model Mean
For the “true” probability, $p_T$ to be unbiased with respect to $p_m$, it is necessary
that
\[ E(p_T) = E\left(\frac{e^Z}{1 + e^Z}\right) = p_m. \] (6.24)

Set \( f(Z, p_m) = \frac{e^Z}{1 + e^Z} \) and \( m' = \text{logit}(p_m) + m \), then \( E(p_T) \) may be approximated using the second order Taylor series

\[
\begin{align*}
    f(Z, p_m) &\approx f(m', p_m) + \left. \frac{\partial f}{\partial Z} \right|_{m'} (Z - m') + \frac{1}{2} \left. \frac{\partial^2 f}{\partial Z^2} \right|_{m'} (Z - m')^2 \\
    &= \frac{e^{m'}}{1 + e^{m'}} + \frac{e^{m'}}{(1 + e^{m'})^2} (Z - m') + \frac{e^{m'} - e^{2m'}}{2(1 + e^{m'})^3} (Z - m')^2. 
\end{align*}
\] (6.25)

Then \( E(p_T) \) is given by
\[
E(p_T) \approx \frac{e^{m'}}{1 + e^{m'}} + \frac{e^{m'} - e^{2m'}}{2(1 + e^{m'})^3} \text{Var}(Z) \\
= p_m 
\] (6.26)

Equation 6.26 may be solved for \( m \) by minimising
\[
\left\| \frac{e^{m'}}{1 + e^{m'}} + \frac{e^{m'} - e^{2m'}}{2(1 + e^{m'})^3} \sigma^2 - p_m \right\| 
\] (6.27)
using the Matlab function \texttt{fminbnd}.

**Simulation of Outcomes**

For simulation of the “true” outcomes \( Y \) using the model given by Equation (6.23), \( \text{logit}(p_T) \) is drawn from an \( N(m', \sigma^2) \) distribution; the “true” probability of mortality calculated as
\[ p_T = \frac{e^{\text{logit}(p_T)}}{1 + e^{\text{logit}(p_T)}}, \] (6.28)
and the outcomes \( Y \) generated as described in Section 6.2. For this model the “true” probability of mortality, \( p_T \), is unbiased with respect to \( p_m \).

### 6.3 Results from ARL Approximations with Patient Level Uncertainty in Risk Adjustment

The results of simulation studies, that use the algorithms outlined in Section 6.2 to approximate the ARL\(_\alpha\)s of CUSUM schemes to monitor mortality of patient populations, are presented in this section. The respective simulation algorithms proposed in Sections 6.2.1, 6.2.2, 6.2.3, 6.2.4, and 6.2.5 will be called the Beta
model, the Gamma model, the Gamma correction, the Normal model, and the Normal correction.

Each algorithm was used to approximate the ARL$_0$s of the risk adjusted and minimum effect CUSUM schemes, defined in Chapters 5.3.2 and 5.5, respectively. ARL$_0$s were found for each of the CUSUM schemes, with decision intervals, $h$, set at 2.5, 3.5, and 4.5, monitoring five cardiac surgical populations risk adjusted by the EuroSCORE, five cardiac surgical populations risk adjusted by the Parsonnet score, and one ICU population risk adjusted by the APACHE III score, where the respective means of the predicted probabilities of mortality for these populations were given in Chapter 5.3.1 as 0.007, 0.009, 0.017, 0.020, 0.024; 0.030, 0.034, 0.046, 0.051, 0.058; and 0.160. Approximate ARL$_0$s were found for the standard deviation, $\sigma$, as defined in Section 6.2, set at the low, high, and very high levels of 0.45, 0.90, and 1.80. There was a total of 198 ARL$_0$s approximated using the simulation algorithms given in Section 6.2. Each ARL$_0$ was estimated from the average of 10,000 simulated run lengths.

The results of each of the simulation studies to approximate ARL$_0$s of the risk adjusted or minimum effect CUSUM schemes using one of the five algorithms are presented as plots in a chart. Each chart has four sub-charts, and each sub-chart contains three plots in which approximate ARL$_0$s found for a CUSUM scheme with $h$ set at 2.5, 3.5, or 4.5 are plotted against the mean predicted probability of mortality, $\bar{p}$, of the eleven patient populations being monitored. The ARL$_0$s in Sub-chart (a) are those given for the risk adjusted CUSUM scheme in Figure 5.6 or for the minimum effect CUSUM scheme in Table 5.4. They are assumed to be the “true” ARL$_0$s, because they were approximated using a process in which deaths were simulated using probabilities of mortality, $p_m$, calculated from the risk model, and are used as benchmarks against which the ARL$_0$s in Sub-charts (b), (c), and (d) are compared. The ARL$_0$s found with the standard deviation, $\sigma$, set at 0.45, 0.90, and 1.80 are given in Sub-charts (b), (c), and (d), respectively.

There are also four tables interspersed among the charts. Each table lists the ratios of ARL$_0$s in Sub-charts (b), (c), and (d) to their benchmarks in Sub-chart (a), where the ARL$_0$s are the approximations for minimum effect and risk adjusted CUSUM schemes with $h = 4.5$ found using the Gamma model, Gamma
correction, Normal model, or Normal correction. The purpose of these tables is to clearly identify where the approximate $\text{ARL}_0$s, found using a process in which patient deaths are simulated from uncertain probabilities of mortality, $p_T$, are more than 5% above or below the approximations found using a process in which patient deaths are simulated from probabilities of mortality, $p_m$, predicted by the risk score.

Beta Model
The approximate $\text{ARL}_0$s of minimum effect CUSUM schemes, where the $\text{ARL}_0$s were estimated using the Beta model, are given in Figure 6.1 and those of risk adjusted schemes are in Figure 6.2. The $\text{ARL}_0$s in the four sub-charts in both figures are similar indicating that the approximations of the $\text{ARL}_0$s of both the risk adjusted and minimum effect CUSUM schemes are unaffected by the uncertainty in the true probability of a patient’s mortality, $p_T$, if it is such that $E(p_T) = p_m$, where $p_m$ is the expected probability that the patient will die as predicted by the risk score.

Gamma Model
Figures 6.3 and 6.4 give approximate $\text{ARL}_0$s simulated using the Gamma model, where the approximations in Figure 6.3 are for minimum effect CUSUM schemes and those in Figure 6.4 are for risk adjusted CUSUM schemes. The plots in Sub-charts (b), (c), and (d) in both figures show that $\text{ARL}_0$s approximated using the Gamma model tend to be greater than their benchmarks in Sub-chart (a). The upward bias in the approximations increases as either the variability of the true odds of patient mortality $\text{Var}(\text{Odds}_T)$, the average of the expected probability of patient mortality $\bar{p}$, or the decision interval $h$ of the CUSUM scheme increases.

Table 6.1 lists ratios of approximate $\text{ARL}_0$s and their benchmarks, where the approximations for the minimum effect and risk adjusted CUSUM schemes with $h = 4.5$ were found using the Gamma model. The $\text{ARL}_0$s of the minimum effect schemes exceed their benchmarks by more than 5% for the three populations with their mean expected probabilities of mortality $\bar{p} \geq 0.051$ if $\sigma = 0.45$, for the eight populations with $\bar{p} \geq 0.020$ if $\sigma = 0.90$, and for the ten populations with $\bar{p} \geq 0.009$ if $\sigma = 1.80$. On the other hand, all the approximate $\text{ARL}_0$s of
Figure 6.1: Approximate ARL$_0$s of minimum effect CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the “true” probability of mortality, $p_T$, drawn from a Beta($\alpha, \beta$) distribution such that $E(p_T) = p_m$. 
Figure 6.2: Approximate ARL₀s of risk adjusted CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the “true” probability of mortality, $p_T$, drawn from a Beta($\alpha, \beta$) distribution such that $E(p_T) = p_m$. 

(a) No Added Noise

(b) Std Dev of Noise = 0.45

(c) Std Dev of Noise = 0.9

(d) Std Dev of Noise = 1.8
Figure 6.3: Approximate $\text{ARL}_0$ of minimum effect CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the “true” odds of mortality, $\text{Odds}_T$, drawn from a $\text{Gamma}(\alpha, \beta)$ distribution such that $E(\text{Odds}_T) = \text{Odds}_m$, the predicted odds of mortality. The maximum values of the scales on the $\text{ARL}_0$ axes increase as the standard deviation increases. The maximums are 5,000, 5,000, 8,000, and 22,000 in Sub-charts (a), (b), (c), and (d), respectively.
Figure 6.4: Approximate $ARL_0$s of risk adjusted CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the “true” odds of mortality, Odds$_T$, drawn from a Gamma$(\alpha, \beta)$ distribution such that $E(\text{Odds}_T) = \text{Odds}_m$, the predicted odds of mortality. The maximum values of the scales on the $ARL_0$ axes increase as the standard deviation increases. The maximums are 50,000, 50,000, 56,000, and 700,000 in Sub-charts (a), (b), (c), and (d), respectively.
Table 6.1: Ratios of ARL0s approximated from the Gamma (Γ) model given in Section 6.2.2 and ARL0s found using the approximation process described in Chapter 5. Approximate ARL0s were estimated for minimum effect and risk adjusted CUSUM schemes with h = 4.5 monitoring eleven populations with varying mean expected probability of mortality (\(\bar{p}\)) and with three levels of uncertainty (\(\sigma\)) for ARL0s found using the Γ model.

<table>
<thead>
<tr>
<th>(\bar{p})</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.007</td>
<td>0.99</td>
<td>1.01</td>
<td>1.02</td>
<td>1.02</td>
<td>1.09</td>
<td>1.41</td>
</tr>
<tr>
<td>0.009</td>
<td>1.02</td>
<td>1.02</td>
<td>1.08</td>
<td>1.03</td>
<td>1.17</td>
<td>1.66</td>
</tr>
<tr>
<td>0.017</td>
<td>1.03</td>
<td>1.04</td>
<td>1.18</td>
<td>1.09</td>
<td>1.31</td>
<td>2.55</td>
</tr>
<tr>
<td>0.020</td>
<td>1.00</td>
<td>1.06</td>
<td>1.21</td>
<td>1.10</td>
<td>1.42</td>
<td>2.98</td>
</tr>
<tr>
<td>0.024</td>
<td>1.04</td>
<td>1.08</td>
<td>1.28</td>
<td>1.12</td>
<td>1.51</td>
<td>3.59</td>
</tr>
<tr>
<td>0.030</td>
<td>1.05</td>
<td>1.09</td>
<td>1.49</td>
<td>1.06</td>
<td>1.19</td>
<td>2.03</td>
</tr>
<tr>
<td>0.034</td>
<td>1.02</td>
<td>1.12</td>
<td>1.62</td>
<td>1.06</td>
<td>1.24</td>
<td>2.30</td>
</tr>
<tr>
<td>0.046</td>
<td>1.02</td>
<td>1.17</td>
<td>1.90</td>
<td>1.10</td>
<td>1.45</td>
<td>3.58</td>
</tr>
<tr>
<td>0.051</td>
<td>1.07</td>
<td>1.26</td>
<td>2.20</td>
<td>1.08</td>
<td>1.47</td>
<td>4.11</td>
</tr>
<tr>
<td>0.058</td>
<td>1.08</td>
<td>1.30</td>
<td>2.56</td>
<td>1.13</td>
<td>1.58</td>
<td>4.65</td>
</tr>
<tr>
<td>0.160</td>
<td>1.24</td>
<td>2.11</td>
<td>7.10</td>
<td>1.54</td>
<td>5.26</td>
<td>126</td>
</tr>
</tbody>
</table>
Figure 6.5: Approximate ARL$_0$s of minimum effect CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the “true” odds of mortality, Odds$_T$, drawn from a Gamma($\alpha, \beta$) distribution, with mean chosen such that $E(p_T) \approx p_m$. 

(a) No Added Noise

(b) Std Dev of Noise = 0.45

(c) Std Dev of Noise = 0.9

(d) Std Dev of Noise = 1.8

Average Prob. of Mortality

Average Prob. of Mortality

ARL

ARL

$0$ $0$ $2,500$ $5,000$ $2,500$ $5,000$ $0.00$ $0.05$ $0.10$ $0.15$ $0.00$ $0.05$ $0.10$ $0.15$
SECTION 3: PATIENT LEVEL UNCERTAINTY

Figure 6.6: Approximate ARL_{0.05} of risk adjusted CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, p_{m}, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the “true” odds of mortality, Odds_{T}, drawn from a Gamma(\alpha, \beta) distribution, with mean chosen such that E(p_T) \approx p_{m}.
Table 6.2: Ratios of $\text{ARL}_0$s approximated using the Gamma correction ($\Gamma$ correct) model given in Section 6.2.3 and $\text{ARL}_0$s found using the approximation process described in Chapter 5. Approximate $\text{ARL}_0$s were estimated for minimum effect and risk adjusted CUSUM schemes with $h = 4.5$ monitoring eleven populations with varying mean expected probability of mortality ($\bar{\rho}$) and with three levels of uncertainty ($\sigma$) for $\text{ARL}_0$s found using the $\Gamma$ correction.

<table>
<thead>
<tr>
<th>$\bar{\rho}$</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.007</td>
<td>1.01</td>
<td>1.00</td>
<td>0.99</td>
<td>1.01</td>
<td>1.00</td>
<td>0.96</td>
</tr>
<tr>
<td>0.009</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>1.00</td>
<td>1.00</td>
<td>0.90</td>
</tr>
<tr>
<td>0.017</td>
<td>1.00</td>
<td>0.98</td>
<td>0.98</td>
<td>1.00</td>
<td>0.96</td>
<td>0.82</td>
</tr>
<tr>
<td>0.020</td>
<td>0.98</td>
<td>1.01</td>
<td>0.98</td>
<td>0.99</td>
<td>0.96</td>
<td>0.77</td>
</tr>
<tr>
<td>0.024</td>
<td>1.00</td>
<td>0.99</td>
<td>0.96</td>
<td>0.98</td>
<td>0.97</td>
<td>0.76</td>
</tr>
<tr>
<td>0.030</td>
<td>0.99</td>
<td>0.99</td>
<td>0.94</td>
<td>1.01</td>
<td>0.98</td>
<td>0.91</td>
</tr>
<tr>
<td>0.034</td>
<td>0.99</td>
<td>0.99</td>
<td>0.92</td>
<td>1.00</td>
<td>0.98</td>
<td>0.86</td>
</tr>
<tr>
<td>0.046</td>
<td>0.98</td>
<td>0.96</td>
<td>0.86</td>
<td>0.97</td>
<td>0.97</td>
<td>0.75</td>
</tr>
<tr>
<td>0.051</td>
<td>1.00</td>
<td>0.99</td>
<td>0.84</td>
<td>0.98</td>
<td>0.95</td>
<td>0.72</td>
</tr>
<tr>
<td>0.058</td>
<td>1.00</td>
<td>0.96</td>
<td>0.80</td>
<td>1.00</td>
<td>0.94</td>
<td>0.70</td>
</tr>
<tr>
<td>0.160</td>
<td>1.01</td>
<td>0.93</td>
<td>0.74</td>
<td>0.99</td>
<td>0.89</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Figure 6.7: Approximate ARL\(_0\)s of minimum effect CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, \(p_m\), predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the logit of the “true” probability of mortality, \(\logit(p_T)\), drawn from an \(N(\logit(p_m), \sigma^2)\) distribution such that \(E(\logit(p_T)) = \logit(p_m)\), the logit of the predicted probability of mortality. The maximum values of the scales on the ARL\(_0\) axes decrease as the standard deviation increases. The maximums are 5,000, 4,000, 2,000, and 250 in Sub-charts (a), (b), (c), and (d), respectively.
Figure 6.8: Approximate ARL₀ values of risk adjusted CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, \( p_m \), predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the logit of the “true” probability of mortality, \( \text{logit}(p_T) \), drawn from an \( N(\text{logit}(p_m), \sigma^2) \) distribution such that \( \text{E}(\text{logit}(p_T)) = \text{logit}(p_m) \), the logit of the predicted probability of mortality. The maximum values of the scales on the ARL₀ axes decrease as the standard deviation increases. The maximums are 50,000, 25,000, 4,000, and 400 in Sub-charts (a), (b), (c), and (d), respectively.
Table 6.3: Ratios of ARL₀s approximated using the Normal (N) model given in Section 6.2.4 and ARL₀s found using the approximation process described in Chapter 5. Approximate ARL₀s were estimated for minimum effect and risk adjusted CUSUM schemes with h = 4.5 monitoring eleven populations with varying mean expected probability of mortality (\(\bar{p}\)) and with three levels of uncertainty (\(\sigma\)) for ARL₀s found using the N model.

<table>
<thead>
<tr>
<th>(\bar{p})</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.007</td>
<td>0.80</td>
<td>0.41</td>
<td>0.05</td>
<td>0.48</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>0.009</td>
<td>0.79</td>
<td>0.39</td>
<td>0.05</td>
<td>0.49</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>0.017</td>
<td>0.75</td>
<td>0.33</td>
<td>0.04</td>
<td>0.50</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>0.020</td>
<td>0.75</td>
<td>0.32</td>
<td>0.04</td>
<td>0.51</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>0.024</td>
<td>0.73</td>
<td>0.31</td>
<td>0.04</td>
<td>0.51</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>0.030</td>
<td>0.66</td>
<td>0.22</td>
<td>0.03</td>
<td>0.48</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>0.034</td>
<td>0.65</td>
<td>0.21</td>
<td>0.03</td>
<td>0.50</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>0.046</td>
<td>0.61</td>
<td>0.20</td>
<td>0.03</td>
<td>0.51</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>0.051</td>
<td>0.62</td>
<td>0.20</td>
<td>0.03</td>
<td>0.51</td>
<td>0.11</td>
<td>0.02</td>
</tr>
<tr>
<td>0.058</td>
<td>0.62</td>
<td>0.19</td>
<td>0.03</td>
<td>0.52</td>
<td>0.12</td>
<td>0.02</td>
</tr>
<tr>
<td>0.160</td>
<td>0.73</td>
<td>0.31</td>
<td>0.05</td>
<td>0.68</td>
<td>0.27</td>
<td>0.05</td>
</tr>
</tbody>
</table>
Figure 6.9: Approximate ARL0s of minimum effect CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the logit of the “true” probability of mortality, $\text{logit}(p_T)$, drawn from a $N(m, \sigma^2)$ distribution, with mean, $m$, chosen such that $E(p_T) \approx p_m$. The maximum value of the scale on the ARL0 axis in Sub-chart (d) is less than the maximums in the other sub-charts. The maximums are 2,000 in Sub-chart (d) and 5,000 in the others.
Figure 6.10: Approximate ARL$_0$s of risk adjusted CUSUM schemes monitoring mortality of eleven patient populations, described in Chapter 5.3.1. In Plot (a), deaths were simulated using the expected probability of mortality, $p_m$, predicted by the risk score and, in Plots (b), (c), and (d), deaths were simulated from the logit of the “true” probability of mortality, logit($p_T$), drawn from a $N(m, \sigma^2)$ distribution, with mean, $m$, chosen such that $E(p_T) \approx p_m$. The maximum values of the scales on the ARL$_0$ axes decrease as the standard deviation increases. The maximums are 50,000, 50,000, 30,000, and 4,000 in Sub-charts (a), (b), (c), and (d), respectively.
Table 6.4: Ratios of ARL₀s approximated using the Normal correction (N correct) model given in Section 6.2.5 and ARL₀s found using the approximation process described in Chapter 5. Approximate ARL₀s are estimated for minimum effect and risk adjusted CUSUM schemes with \( h = 4.5 \) monitoring eleven populations with varying mean expected probability of mortality (\( \bar{p} \)) and with three levels of uncertainty (\( \sigma \)) for ARL₀s found using the N correct model.

<table>
<thead>
<tr>
<th>( \bar{p} )</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
<th>0.45</th>
<th>0.90</th>
<th>1.80</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.007</td>
<td>0.98</td>
<td>0.88</td>
<td>0.31</td>
<td>0.96</td>
<td>0.66</td>
<td>0.06</td>
</tr>
<tr>
<td>0.009</td>
<td>1.01</td>
<td>0.87</td>
<td>0.30</td>
<td>0.98</td>
<td>0.69</td>
<td>0.07</td>
</tr>
<tr>
<td>0.017</td>
<td>1.00</td>
<td>0.86</td>
<td>0.28</td>
<td>0.99</td>
<td>0.76</td>
<td>0.10</td>
</tr>
<tr>
<td>0.020</td>
<td>0.99</td>
<td>0.87</td>
<td>0.28</td>
<td>0.98</td>
<td>0.78</td>
<td>0.12</td>
</tr>
<tr>
<td>0.024</td>
<td>0.99</td>
<td>0.87</td>
<td>0.29</td>
<td>0.99</td>
<td>0.79</td>
<td>0.14</td>
</tr>
<tr>
<td>0.030</td>
<td>1.00</td>
<td>0.83</td>
<td>0.25</td>
<td>0.98</td>
<td>0.74</td>
<td>0.11</td>
</tr>
<tr>
<td>0.034</td>
<td>1.00</td>
<td>0.84</td>
<td>0.26</td>
<td>0.97</td>
<td>0.76</td>
<td>0.13</td>
</tr>
<tr>
<td>0.046</td>
<td>0.96</td>
<td>0.85</td>
<td>0.29</td>
<td>1.00</td>
<td>0.82</td>
<td>0.21</td>
</tr>
<tr>
<td>0.051</td>
<td>0.99</td>
<td>0.87</td>
<td>0.31</td>
<td>1.00</td>
<td>0.84</td>
<td>0.23</td>
</tr>
<tr>
<td>0.058</td>
<td>1.00</td>
<td>0.89</td>
<td>0.35</td>
<td>0.98</td>
<td>0.88</td>
<td>0.24</td>
</tr>
<tr>
<td>0.160</td>
<td>1.00</td>
<td>0.95</td>
<td>0.51</td>
<td>1.00</td>
<td>1.01</td>
<td>0.81</td>
</tr>
</tbody>
</table>
the risk adjusted CUSUM scheme with $h = 4.5$ exceed their benchmarks by more than 5% except, where $\sigma = 0.45$, the ARL$_0$s of the risk adjusted schemes monitoring the populations with $\bar{p} = 0.007$ and 0.009. The ratios in each column of Table 6.1, which are for ARL$_0$s approximated with constant decision interval, $h$, and standard deviation, $\sigma$, increase as the mean of the predicted probability of patient mortality, $\bar{p}$, increases. That is, ARL$_0$s approximated using the Gamma model appear to be related to the average predicted probability of mortality of the population being monitored. However, the monotone increases of the ratios of the minimum effect schemes may be contrasted with the piecewise increases of the risk adjusted schemes, which indicate that ratios of risk adjusted CUSUM schemes are also related to the risk model used to predict probabilities of patient mortality.

**Gamma Correction**

Comparison of the respective plots of the approximate ARL$_0$s of minimum effect and risk adjusted CUSUM schemes in Figures 6.5 (b) and 6.6 (b) with their counterparts in the Sub-charts (a) show no differences between ARL$_0$s approximated using the Gamma correction with $\sigma = 0.45$ and their benchmarks. The ARL$_0$s for the minimum effect scheme with $h = 4.5$ in Figure 6.5 (c), where $\sigma = 0.90$, and those for the CUSUM schemes with $h = 3.5$ and 4.5 in Sub-chart (d), where $\sigma = 1.80$, show a downward trend, in comparison with their benchmarks, as the mean probability of mortality, $\bar{p}$, of the patient population increases. Similarly, the overcorrection of the ARL$_0$ approximations, in Figure 6.6 (c) for the risk adjusted schemes with $h = 4.5$ and in Figure 6.6 (d) for the risk adjusted schemes with $h = 3.5$ and 4.5, increase as $\bar{p}$ increases.

Table 6.2 gives the ratios of ARL$_0$s approximated using the Gamma correction and their benchmarks. In the list for minimum effect schemes where the approximations were found with $\sigma = 0.90$, the only ARL$_0$ underestimated is that of the minimum effect scheme monitoring ICU patient populations with $\bar{p} = 0.160$. If $\sigma = 1.80$, the approximate ARL$_0$s are less than 95% their benchmarks for the minimum effect schemes monitoring the seven populations with $\bar{p} \geq 0.030$. For the risk adjusted schemes, the approximate ARL$_0$s, found using the Gamma correction with $\sigma = 0.90$, are $\leq 95\%$ for the three patient populations with $\bar{p} \geq 0.051$ and,
for $\sigma = 1.80$, only the ratio for the approximate $\text{ARL}_0$ of the risk adjusted scheme monitoring the population with $\bar{p} = 0.007$ is greater than 0.95. If the ratios for both the risk adjusted and minimum effect CUSUM schemes monitoring a patient population are less than 0.95, the ratio for the risk adjusted scheme tends to be the lower. The ratios of the approximate $\text{ARL}_0$s of both CUSUM schemes tend to decrease as $\bar{p}$ increases, but the decrease with increasing $\bar{p}$ is monotone for the ratios of the $\text{ARL}_0$s of the minimum effect scheme and piecewise for those of the risk adjusted scheme. There appears to be a relationship between the downward trend of the ratios of the $\text{ARL}_0$s of risk adjusted CUSUM schemes and the risk models that is comparable with that seen in the piecewise increases of ratios of the approximate $\text{ARL}_0$s of risk adjusted schemes given in Table 6.1, where the $\text{ARL}_0$s were approximated using the Gamma model.

Normal Model
Approximate $\text{ARL}_0$s simulated using the Normal model are presented in Figures 6.7 and 6.8. The $\text{ARL}_0$s for minimum effect CUSUM schemes in Figures 6.7 (b), (c), and (d) clearly show that the approximations found using the Normal model are less than their benchmarks given in Sub-chart (a) for all values of $\sigma$. The approximate $\text{ARL}_0$s decrease as $\sigma$ increases and, for $\sigma = 1.80$, they are comparable with the “true” $\text{ARL}_1$s given in Table 5.5 for minimum effect schemes monitoring out-of-control processes. That is, if there is a very high level of uncertainty in the procedure outlined in Section 6.2.4, the effect of the upward bias of $p_T$ with respect to $p_m$ is so great that the magnitude of the approximate $\text{ARL}_0$s of risk adjusted CUSUM schemes found using this procedure is indistinguishable from that of the out-of-control $\text{ARL}_0$s estimated using the procedure given in Section 5.3.3, where it is assumed that $p_m = p_T$. Similarly, the approximate $\text{ARL}_0$s of risk adjusted schemes, given in Figure 6.8, decrease as $\sigma$ increases. The approximations in Sub-chart (d), found with $\sigma = 1.80$, are also comparable with the “true” $\text{ARL}_1$s given in Figure 5.7 for risk adjusted CUSUM schemes monitoring out-of-control processes. The $\text{ARL}_0$ performance measures approximated using the normal (logistic) model provided some evidence that, if the downward bias of $p_m$ with respect to $p_T$ is sufficiently large, risk adjusted CUSUM monitoring schemes will be unable to
distinguish a process that is in statistical control from one that is out of statistical control.

The ratios in Table 6.3 highlight the degree of underestimation of ARLₐ₀s approximated using the Normal model. It is best illustrated by the approximations of the ARLₐ₀s listed in the columns where σ = 1.80. For the minimum effect CUSUM scheme, they are ≤ 5% of their “true” values for all populations monitored. If monitoring is by the risk adjusted CUSUM scheme, they are one or two percent of their “true” values for all but the ICU population, where the ratio is 0.05. Paradoxically, if the Normal model is used to approximate the ARLₐ₀s of minimum effect and risk adjusted CUSUM schemes, the apparent relationship between the risk models used to predict the probability of patient mortality and the ratios in Table 6.3 is for the ratios of the ARLₐ₀s of the minimum effect schemes rather than the ratios of the ARLₐ₀s of risk adjusted schemes, as seen in Tables 6.1 and 6.2, where the approximate ARLₐ₀s were found using the Gamma model and Gamma correction, respectively. For approximations using the Normal model with σ = 0.45, the ratios of the ARLₐ₀s of minimum effect schemes are in the intervals (0.73, 0.80) and (0.62, 0.66) for the cardiac surgical populations risk adjusted by the EuroSCORE and the Parsonnet score, respectively, and the ratio for the ICU population risk adjusted by the APACHE III score is 0.73. The ratios of the ARLₐ₀s of minimum effect schemes approximated with σ = 0.90 are less than those approximated with σ = 0.45, but both sets of ratios have a similar pattern of higher for the EuroSCORE, lower for the Parsonnet score, and higher for the APACHE III score.

**Normal Correction**

Comparison of the approximate ARLₐ₀s plotted in Figures 6.9 (b) and 6.10 (b) with their benchmarks in Sub-chart (a) show that, if σ = 0.45, the Normal correction eliminates the downward bias in the approximations found using the Normal model. However, the approximate ARLₐ₀s of the minimum effect CUSUM schemes given in Figure 6.9 (c), where σ = 0.90, are somewhat less than their benchmarks. The lowest ARLₐ₀s for each of the minimum effect schemes, where h = 2.5, 3.5, or 4.5, occur for the cardiac surgical populations risk adjusted by Parsonnet score. In
Figure 6.9 (d), where $\sigma = 1.80$, the approximations for the minimum effect schemes monitoring cardiac surgical populations are less than half their counterparts in Figure 6.9 (a). The lowest $ARL_0$ occurs for each of the schemes with $h = 2.5, 3.5,$ or $4.5$ monitoring the population with $\bar{p} = 0.030$.

For risk adjusted schemes, the approximate $ARL_0$s in Figures 6.10 (c) and (d) are also less than their counterparts in Figures 6.10 (a). In Sub-chart (c), where $\sigma = 0.90$, the $ARL_0$s of each of the schemes with $h = 2.5, 3.5,$ or $4.5$ decrease from a maximum at the minimum $\bar{p} = 0.007$ to a minimum at the maximum $\bar{p} = 0.160$, as the benchmark $ARL_0$s in Sub-chart (a) do. However, the plots in Figure 6.10 (d), where $\sigma = 1.80$, shows that the minimum $ARL_0$s of each of the three risk adjusted CUSUM schemes occur at $\bar{p} = 0.030$ and the maximum $ARL_0$ of the risk adjusted schemes with $h = 4.5$ is at $\bar{p} = 0.160$.

Table 6.4 lists the ratios of the $ARL_0$s of both the minimum effect and risk adjusted CUSUM schemes to their benchmarks for $ARL_0$s approximated using the Normal correction. The ratios of the $ARL_0$s of the minimum effect schemes are approximately one for $\sigma = 0.45$, but, for the schemes monitoring cardiac surgical populations, they somewhat less than one for $\sigma = 0.90$ and less than 0.5 for $\sigma = 1.80$. The ratios of the $ARL_0$s of minimum effect schemes monitoring ICU populations with $\bar{p} = 0.160$ are also less than one for the approximations found using the Normal correction with $\sigma = 0.90$ and 1.80, but they are greater than the $ARL_0$s for minimum effect CUSUM schemes monitoring cardiac surgical populations. The ratios of the $ARL_0$s of risk adjusted schemes are, in general, less than one for $ARL_0$ approximations found using the Normal correction with $\sigma = 0.90$ or 1.80. In general, they decrease as $\sigma$ increases and increase as $\bar{p}$ increases. In contrast to the apparent relationship between the risk models and the ratios of approximations found using the Normal model, it is the ratios of the $ARL_0$s of risk adjusted rather than minimum effect CUSUM schemes that give some indication of a relationship with the risk scores. If $\sigma = 1.80$, all the ratios for risk adjusted CUSUM schemes monitoring cardiac surgical populations were less than the equivalent ratios for minimum effect schemes, but, surprisingly, the ratio of $ARL_0$s for risk adjusted CUSUM scheme monitoring the ICU patients, where $\bar{p} = 0.160$, is much higher than the ratio for the minimum effect scheme monitoring
the same patients.

**Comparisons of ARL\(_0\)s with \(\bar{p}\) Fixed**

The approximate ARL\(_0\)s of risk adjusted and minimum effect CUSUM schemes monitoring patient populations with \(\bar{p} = 0.020, 0.051,\) and 0.160 are presented in another way in Figures 6.11, 6.12, and 6.13, respectively. Each figure contains five charts in which the benchmark and approximate ARL\(_0\)s of both the risk adjusted and minimum effect CUSUM schemes are plotted against \(\sigma = 0\) (for the benchmarks), 0.45, 0.90, and 1.80. Charts (a), (b), (c), (d), and (e) present approximations found using the Normal model, the Gamma model, the Beta model, the Normal correction, respectively.

The purpose is to facilitate comparison of the approximate ARL\(_0\)s found using each simulation algorithm. The charts in each figure show that the ARL\(_0\)s from the Normal model rapidly decrease and those from the Gamma model rapidly increase as \(\sigma\) increases, but the approximations found using the Beta distribution are stable and independent of \(\sigma\). The plots for the Normal and Gamma corrections show that, if \(\sigma = 0.45\), their approximate ARL\(_0\)s are similar to their benchmarks. However, for higher values of \(\sigma\), the Normal correction tends to underestimate the ARL\(_0\)s although the underestimation is less than that of the Normal model. The Gamma correction with high values of \(\sigma\) also tends to underestimate the ARL\(_0\)s of both CUSUM schemes, but the approximations are closer to the benchmarks than the overestimates from the Gamma model.

Comparison of the plots in each figure shows the relationship between the approximate ARL\(_0\)s and \(\bar{p}\) of the patient population varies according the method used to approximate the ARL\(_0\)s. For ARL\(_0\)s found using the Normal model, the approximations in Figure 6.13 (a), where \(\bar{p} = 0.160\), decrease a little less rapidly as \(\sigma\) increases than they do in Figures 6.11 (a) and 6.12 (a), where \(\bar{p} = 0.020\) and 0.051, respectively. On the other hand, the ARL\(_0\)s in Figures 6.11 (b), 6.12 (b), and 6.13 (b) show that, for approximations found using the Gamma model, the magnitude of their increase above their benchmarks is least for both the risk adjusted and minimum effect schemes if \(\bar{p} = 0.020\) and greatest if \(\bar{p} = 0.160\). The Normal and Gamma corrections are designed to simulate ARL\(_0\) approximations that
CHAPTER 6: RA-CUSUM—PREDICTION UNCERTAIN

Figure 6.11: Approximate $ARL_0$'s of risk adjusted and minimum effect CUSUM schemes monitoring deaths in a patient population with mean expected probability of mortality $\bar{p} = 0.020$. Each plot shows the relationship of the $ARL_0$ approximations with increasing uncertainty in the “true” probability of patient mortality, where it is modelled using one of the five methods outlined in Section 6.2. The maximum value of the scale on the $ARL_0$ axis is 60,000 for Sub-chart (b) and 20,000 all other sub-charts.
Figure 6.12: Approximate $\text{ARL}_0$s of risk adjusted and minimum effect CUSUM schemes monitoring deaths in a patient population with mean expected probability of mortality $\bar{p} = 0.051$. Each plot shows the relationship of the $\text{ARL}_0$ approximations with increasing uncertainty in the “true” probability of patient mortality, where it is modelled using one of the five methods outlined in Section 6.2. The maximum value of the scale on the $\text{ARL}_0$ axis is 32,000 for Sub-chart (b) and 8,000 all other sub-charts.
Figure 6.13: Approximate $A_{RL0}$s of risk adjusted and minimum effect CUSUM schemes monitoring deaths in a patient population with mean expected probability of mortality $\bar{\rho} = 0.160$. Each plot shows the relationship of the $A_{RL0}$ approximations with increasing uncertainty in the “true” probability of patient mortality, where it is modelled using one of the five methods outlined in Section 6.2. The maximum value of the scale on the $A_{RL0}$ axis is 690,000 for Sub-chart (b) and 6,000 all other sub-charts.
are more stable with respect to their benchmarks than the ARL_0's simulated using the Normal and Gamma models are. Comparison of the plots in Figures 6.11 (d), 6.12 (d), and 6.13 (d) show that the ARL_0's approximated using the Normal correction are most stable for CUSUM schemes monitoring populations with \( \bar{p} = 0.160 \), but Figures 6.11 (e), 6.12 (e), and 6.13 (e) show that the ARL_0 approximations found using the Gamma correction are most stable for schemes monitoring populations with \( \bar{p} = 0.020 \). That is, it appears that, if approximate ARL_0's are found using the Normal correction, their stability increases as \( \bar{p} \) of the population being monitored increases, but, if they are found using the Gamma correction, their stability decreases as \( \bar{p} \) increases.

6.4 Discussion of Effect of Patient Level Uncertainty on Risk Adjusted CUSUM Performance

The primary purpose for approximating the ARL_0's of the risk adjusted and minimum effect CUSUM schemes using the simulation algorithms described in Section 6.2 was to investigate the effect of uncertainty in the estimation of the probabilities of mortality, \( p_m \), for individual patients on the performance of the CUSUM schemes. The approximate ARL_0's found using the Beta model, where the “true” outcome probabilities, \( p_T \), were randomly generated from a Beta distribution with mean equal to the expected outcome probabilities, \( p_m \), were no different from their benchmark ARL_0's. That is, \( p_T \)s were generated such that \( E(p_T) = p_m \). On the other hand, simulation using the Gamma and Normal models, where the respective \( p_T \)s were generated from Gamma and Normal distributions with means such that \( E(p_T) \neq p_m \), produced ARL_0's which were biased upwards with respect to their benchmarks, if approximated from the Gamma model, and biased downwards, if approximated from the Normal model. The ARL_0's from Gamma and Normal corrections, where the respective \( p_T \)s are drawn from Gamma and Normal distributions with means, \( m_i \), such that \( E(p_T) = p_m \), were unbiased with respect to their benchmarks if the variability of \( p_T \) was low. If the variability of \( p_T \) was high or very high, the approximations were biased with respect to their benchmarks,
but the cause of these biases was attributed to the limitations of the Taylor series approximations used to find \( m \). Overall, these simulation studies indicate that, if \( p_T \) is drawn from any distribution with \( \mathbb{E}(p_T) = p_m \), the approximate ARL\(_0\)s will be unbiased with respect to benchmark ARL\(_0\)s, which were simulated using a process in which it was assumed \( p_T = p_m \). If risk models are well calibrated to the population of interest, \( \mathbb{E}(p_T) = p_m \). Thus, the results given in Section 6.3 have provided evidence that, for well calibrated risk models, the uncertainty in estimating the expected outcomes of individuals will not affect the performance, as assessed by the ARL\(_0\) measure, of risk adjusted and minimum effect CUSUM schemes.

The downward bias of ARL\(_0\)s approximated using the Normal model has implications for the risk scores, such as the EuroSCORE, the Parsonnet score, and the APACHE III score, derived using logistic regression models. For such scores, the predicted probability of mortality, \( p_m \), is given by

\[
\text{logit}(p_m) = X_m \beta,
\]

where it is assumed the measurement of the covariates, \( X_m \) are error free, but some recent publications (Cole et al., 2006; Fraser and Yan, 2007; Qin and Zhang, 2007) have noted that measurement error is ubiquitous in medical studies. Suppose \( X_m \) is measured with error \( \varepsilon \) such that \( \varepsilon \sim N(0, \sigma^2) \), then \( p_m \) is recorded as \( \tilde{p}_m \) given by

\[
\text{logit}(\tilde{p}_m) = (X_m + \varepsilon) \beta.
\]

It is clear from the results of the study, in which ARL\(_0\)s were approximated using the Normal model, that \( \tilde{p}_m \) is biased upwards with respect to \( p_m \) and the amount of bias increases as \( \sigma^2 \) increases.

There are reports of upward bias in models to predict the probability of mortality following cardiac surgery. Bernstein and Parsonnet (2000) note that inclusion of explanatory variables that duplicate or reinforce risk factors already present in the model inevitably results in overestimation of preoperative risk. Measurement error in the predictors provides a possible explanation for upward bias of \( p_m \)s derived using logistic regression. For example, suppose that risk factors that duplicate or reinforce existing explanatory variables are added to a logistic model \( K \) with \( k \).
covariates such that the new model $K'$ has $k'$ covariates. Let the measurement error for each covariate $X_i$, for $i = 1, \ldots, k'$, be $\varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma_i^2)$, then the variances for the error measurement in models $K$ and $K'$ are

$$
\sigma^2_K = \sum_{i=1}^{k} \beta_{K,i}^2 \sigma_i^2 \quad \text{and} \quad \sigma^2_{K'} = \sum_{i=1}^{k'} \beta_{K',i}^2 \sigma_i^2,
$$

respectively. For $\sigma^2_{K'} > \sigma^2_K$, the amount of upward bias of the probabilities derived from model $K'$ is greater than the amount from model $K$.

In a review of the performance of the EuroSCORE, Gogbashian et al. (2004) reported that EuroSCOREs $\leq 6$ overestimated the probability of mortality. It is known (Roques et al., 1999) that the logistic regression coefficients of the EuroSCORE were fitted using data with minimal measurement error. The data were collected using protocols that are unrealistic for routine surveillance and, before fitting the model, they were audited to ensure that the preset targets of completeness and accuracy of $> 99\%$ had been met. When Nashef et al. (1999) proposed the EuroSCORE, they invited other workers to test it in their hospitals, but, if the protocols and audits for the data collected for the follow up studies were relaxed, it is likely that measurement error for the covariates would increase and upward bias for the EuroSCOREs would occur. That is, an increase in measurement error offers a possible explanation for the overestimation that Gogbashian et al. (2004) reported.

Although the simulation studies presented here and in Chapter 5 indicate that the mean expected probability of mortality of the patient population is a very important determinant of the ARL$_0$s of risk adjusted CUSUM schemes, it appears the choice of risk model or the distribution of the patient population has some effect on the ARL$_0$s of both risk adjusted and minimum effect schemes. The results in Section 6.3 gave indications that the risk scores and the ARL$_0$s of risk adjusted CUSUM schemes were related, if the approximations were found using the Gamma model, Gamma correction, or Normal correction, and the risk scores and the ARL$_0$s of minimum effect schemes were related, if the approximations were found using the Normal model. For the simulation studies investigating the effect of uncertainty in the risk of individual mortality on ARL$_0$ approximations, the distributions of the probabilities of mortality of the cardiac surgical populations risk adjusted by
EuroSCORE are discrete with 20 categories, those of the populations risk adjusted by Parsonnet score are discrete with eight to 48 categories, and the distribution of the probabilities of mortality of the ICU population is continuous. It may be that the distribution of the patient population has some secondary influence on the ARL$_0$s of both CUSUM schemes. However, it will be shown in Section 6.5 that the uncertainty in parameter estimation, when the EuroSCORE and Parsonnet score are recalibrated for local populations, is a more important influence on the ARL$_0$s of both the CUSUM schemes.

### 6.5 Effects of Variability in Model Parameter Estimates

In an example to illustrate the characteristics of the risk adjusted CUSUM scheme, Steiner et al. (2000) used a training data set with 2,218 cardiac surgical patients to recalibrate the risk adjustment score by fitting the logistic model

$$\text{logit}(\hat{p}_t) = \hat{\alpha} + \hat{\beta}X_t,$$

(6.29)

where $\hat{p}_t$ is the estimate of the expected probability of postoperative mortality and $X_t$ is the risk adjustment score for patient $t$. Clearly, there is uncertainty in the estimates $(\hat{\alpha}, \hat{\beta})^T$ of the parameters of the recalibration model and the uncertainty increases as the number of cases in the training data set decreases. It gives rise to variability in $\hat{p}$ which, in turn, affects the performance of risk adjusted and minimum effect CUSUM schemes.

The objective in this section is to investigate the effect that the size of training data sets for recalibrating risk scores has on the variability of $\hat{p}_t$ and the performance of risk adjusted and minimum effect CUSUM schemes. In Section 6.5.1, Taylor series approximations are used to calculate variances and means of $\hat{p}$ predicted from unbiased $(\hat{\alpha}, \hat{\beta})^T$ as the variance-covariance of $(\hat{\alpha}, \hat{\beta})^T$ changes. The results of a study on the distribution of the ARL$_0$s of risk adjusted and minimum effect CUSUM schemes, where $\hat{p}_t$ is calculated using a risk model recalibrated using Equation (6.29) with uncertain $(\hat{\alpha}, \hat{\beta})^T$, are given in Section 6.5.2 and, in
Section 6.5.3, plots for a risk adjusted CUSUM scheme to monitor the number of deaths of patients treated in an ICU are presented to show the effect of variability in the estimation of $\hat{p}$ on the time and frequency of alarms.

### 6.5.1 Effect on Estimated Probability of Outcome

In the first study of the variability of parameter estimates, the approximate means $E(\hat{p})$ and variances $V(\hat{p})$ are computed for the expected probability of mortality, $\hat{p}$, estimated by risk models which have been recalibrated by fitting Equation (6.29) to training data sets. Each training data set was constructed by simulating risks scores and outcomes for patients from one of the two populations given in Figures 5.3 (c) and 5.4 (c), where the cardiac surgical patients are classified according to the Parsonnet scores 0–47 and the EuroSCOREs 0–19, respectively.

The minimum number of cases in a training data set is constrained because at least 20 adverse and 20 satisfactory outcomes are required to satisfy the rule of thumb (Peduzzi et al., 1996) that at least ten events of each type are needed for each explanatory variable in a logistic regression model. The reason is that fewer events may result in major problems in the model fit. The mean expected probability of mortality is 0.0457 for the Parsonnet score population in Figure 5.3 (c) and 0.0168 for the EuroSCORE population in Figure 5.4 (c). Therefore, the number of patients required, on average, for 20 cases of mortality is approximately 440 for the Parsonnet score and approximately 1,200 for the EuroSCORE. Accordingly, training data sets with $n_P$ patients, where $n_P \in \{500, 1000, 2000, 4000\}$, or $n_E$ cases, where $n_E \in \{1200, 2000, 4000\}$, were chosen for recalibration of the Parsonnet score or EuroSCORE, respectively.

The distribution of the parameters, $(\hat{\alpha}, \hat{\beta})^T$, estimated from recalibrating the risk scores, is assumed to be bivariate normal with mean $(\alpha, \beta)^T$, taken to be the values $(-3.68, 0.077)^T$ given by Steiner et al. (2000) for the Parsonnet score and $(-5.565, 0.340)^T$ given in Chapter 5.3.1 for the EuroSCORE. The variances, $V(\hat{\alpha})$ and $V(\hat{\beta})$, and covariance $\text{Cov}(\hat{\alpha}, \hat{\beta})$ depend on the sizes of the training data sets. Given a training data set with $n_P$ cases, say, the values of $V(\hat{\alpha})$, $V(\hat{\beta})$, and $\text{Cov}(\hat{\alpha}, \hat{\beta})$ of $(\hat{\alpha}, \hat{\beta})^T$ were calculated using the following process.
Patients, \( t \), for \( t = 1, \ldots , n_P \), with Parsonnet scores, \( X_t \), for \( X_t = 0, \ldots , 48 \), were drawn from a cardiac surgical population, with the distribution of \( X_t \) given in Figure 5.6 (c). Outcomes, \( Y_t \in \{0, 1\} \), were assigned by first calculating the “true” probabilities of mortality, \( p_t \), by substituting \( X_t \) and the values \((-3.68, 0.077)^T\), assumed to be the “true” parameter values, into Equation (6.29) and then drawing events 0 or 1 with probability \( 1 - p_t \) or \( p_t \), respectively. The estimated parameters, \((\hat{\alpha}, \hat{\beta})^T\), and associated variance-covariance matrix, were found by fitting Equation (6.29) to the simulated data \( Y_t \) and \( X_t \). The variances of \( \hat{\alpha} \) and \( \hat{\beta} \) and covariance of \((\hat{\alpha}, \hat{\beta})^T\) in the variance-covariance matrix were recorded. The process was repeated 1,000 times and the respective means of all the recorded variances and covariances were taken to be \( V(\hat{\alpha}), V(\hat{\beta}) \), and \( \text{Cov}(\hat{\alpha}, \hat{\beta}) \) for \((\hat{\alpha}, \hat{\beta})^T\) estimated using a data set with \( n_P \) cases. Construction of each data set and estimation of the logistic regression parameters \((\hat{\alpha}, \hat{\beta})^T\) were undertaken using the functions in the R, Version 2.0.1, statistical package. A similar process was used to find \( V(\hat{\alpha}), V(\hat{\beta}) \), and \( \text{Cov}(\hat{\alpha}, \hat{\beta}) \) of the estimates \((\hat{\alpha}, \hat{\beta})^T\) found after recalibrating the EuroSCORE using training data sets containing \( n_E \) patients.

Substitution of the parameter estimates \((\hat{\alpha}, \hat{\beta})^T\) and the risk score \( X_t \) for patient \( t \) into Equation (6.29) gives an estimate of the expected probability of mortality, \( \hat{p}_t | X_t \), as

\[
\hat{p}_t | X_t = \frac{e^{\hat{\alpha} + \hat{\beta}X_t}}{1 + e^{\hat{\alpha} + \hat{\beta}X_t}}.
\] (6.30)

For parameter estimates \((\hat{\alpha}, \hat{\beta})^T\) close to the true values \((\alpha, \beta)^T\), \( \hat{p}_t | X_t \) in Equation (6.30) may be approximated by the second order Taylor series given by

\[
\hat{p}_t | X_t \approx \hat{p}_0 + f_2(\hat{\alpha} - \alpha) + f_2(\hat{\beta} - \beta) + \frac{1}{2}\{f_{11}(\hat{\alpha} - \alpha)^2 + 2f_{12}(\hat{\alpha} - \alpha)(\hat{\beta} - \beta) + f_{22}(\hat{\beta} - \beta)^2\},
\] (6.31)

where \( \hat{p}_0 = e^{\alpha + \beta X_t}/(1 + e^{\alpha + \beta X_t}) \) and the functions \( f_1 \) and \( f_2 \) are the partial derivatives of \( \hat{p}_t | X_t \) with respect to (wrt) \( \hat{\alpha} \) and \( \hat{\beta} \), respectively, and \( f_{11}, f_{12}, \) and \( f_{22} \) are the second partial derivatives of \( \hat{p}_t | X_t \) wrt to \( \hat{\alpha} \), \( \hat{\alpha} \) and \( \hat{\beta} \), and \( \hat{\beta} \), respectively.
The partial derivatives are evaluated at \((\alpha, \beta)^T\) as

\[
f_1 = \frac{e^{\alpha + \beta X_t}}{(1 + e^{\alpha + \beta X_t})^2}, \quad f_2 = X_t \frac{e^{\alpha + \beta X_t}}{(1 + e^{\alpha + \beta X_t})^2}
\]

\[
f_{11} = \frac{e^{\alpha + \beta X_t}(1 - e^{\alpha + \beta X_t})}{(1 + e^{\alpha + \beta X_t})^3}, \quad f_{12} = X_t \frac{e^{\alpha + \beta X_t}(1 - e^{\alpha + \beta X_t})}{(1 + e^{\alpha + \beta X_t})^3}
\]

\[
f_{22} = X_t^2 \frac{e^{\alpha + \beta X_t}(1 - e^{\alpha + \beta X_t})}{(1 + e^{\alpha + \beta X_t})^3}.
\]

The approximate mean \(E(\hat{p}_t | X_t)\) is given by the second order Taylor series as

\[
E(\hat{p}_t | X_t) \approx \hat{p}_0 + \frac{1}{2} \left\{ f_{11} V(\hat{\alpha}) + 2 f_{12} \text{Cov}(\hat{\alpha}, \hat{\beta}) + f_{22} V(\hat{\beta}) \right\}.
\] (6.32)

The first order Taylor series approximation (Casella and Berger, 1990, Page 329) of the variance \(V(\hat{p}_t | X_t)\) is given by

\[
V(\hat{p}_t | X_t) \approx f_1^2 V(\hat{\alpha}) + f_2^2 V(\hat{\beta}) + 2 f_1 f_2 \text{Cov}(\hat{\alpha}, \hat{\beta}).
\] (6.33)

\(E(\hat{p}_t | X_t)\) and \(V(\hat{p}_t | X_t)\) may be evaluated for all \(X_t\) of both the Parsonnet score and EuroSCORE. Then, the mean \(E(\hat{p})\) estimated for the cardiac surgical population given in Figure 5.3 (c) or 5.4 (c), where \((\hat{\alpha}, \hat{\beta})^T\) was estimated using a training data set with \(n_P\) or \(n_E\) cases, respectively, may be computed as

\[
E(\hat{p}) = \sum_{i=0}^{k} E(\hat{p}_i | X_i) \text{Pr}(X_i = i),
\] (6.34)

where \(k = 47\) for the Parsonnet score or \(19\) for the EuroSCORE and \(\text{Pr}(X_i = i)\) are the probability masses for the classes of cardiac surgical patients given in Figures 5.3 (c) or 5.4 (c), where the patients are categorized according to the Parsonnet score or EuroSCORE, respectively. The variance \(V(\hat{p})\) is computed using the formula given by Gelman et al. (1995, Equation 1.7) as

\[
V(\hat{p}) = E \{V(\hat{p} | X)\} + V \{E(\hat{p} | X)\}
\]

\[
= \sum_{i=0}^{k} V(\hat{p} | X) \text{Pr}(X = i)
\]

\[
+ \sum_{i=0}^{k} \{E(\hat{p} | X)\}^2 \text{Pr}(X = i) - \{E(\hat{p})\}^2.
\] (6.35)

The values of the standard deviations, \(\text{SD}(\hat{\alpha})\) and \(\text{SD}(\hat{\beta})\), and the associated approximations of the means, \(E(\hat{p})\), and standard deviations, \(\text{SD}(\hat{p})\), estimated
Table 6.5: The means, $E(\hat{p})$, and standard deviations, $SD(\hat{p})$, of the expected probabilities of mortality, found using the Taylor series approximations given in Equations (6.32) and (6.33), respectively, after recalibrating the Parsonnet score and the EuroSCORE by fitting Equation (6.29) to data sets with $n$ cases. The respective standard deviations of the parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, are given in Columns $SD(\hat{\alpha})$ and $SD(\hat{\beta})$.

<table>
<thead>
<tr>
<th>Risk Model</th>
<th>$n$</th>
<th>SD($\hat{\alpha}$)</th>
<th>SD($\hat{\beta}$)</th>
<th>$E(\hat{p})$</th>
<th>SD($\hat{p}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parsonnet Score</td>
<td>500</td>
<td>0.347</td>
<td>0.0264</td>
<td>0.0440</td>
<td>0.0384</td>
</tr>
<tr>
<td></td>
<td>1,000</td>
<td>0.240</td>
<td>0.0180</td>
<td>0.0449</td>
<td>0.0362</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>0.168</td>
<td>0.0124</td>
<td>0.0453</td>
<td>0.0352</td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>0.118</td>
<td>0.0087</td>
<td>0.0455</td>
<td>0.0347</td>
</tr>
<tr>
<td>EuroSCORE</td>
<td>1,200</td>
<td>0.463</td>
<td>0.0674</td>
<td>0.0161</td>
<td>0.0317</td>
</tr>
<tr>
<td></td>
<td>2,000</td>
<td>0.354</td>
<td>0.0511</td>
<td>0.0164</td>
<td>0.0311</td>
</tr>
<tr>
<td></td>
<td>4,000</td>
<td>0.247</td>
<td>0.0356</td>
<td>0.0166</td>
<td>0.0306</td>
</tr>
</tbody>
</table>
after recalibration of the Parsonnet score with training data with 500, 1,000, 2,000, and 4,000 cases and recalibration of the EuroSCORE with training data with 1,200, 2,000 and 4,000 cases, are listed in Table 6.5. One result, that is consistent with the assumption that the estimates \((\hat{\alpha}, \hat{\beta})^T\) are from a bivariate normal distribution, is that the values of the variance-covariance of \((\hat{\alpha}, \hat{\beta})^T\) double if number of cases in the training data set halves. However, the variability of the parameter estimates is considerable. For example, if a training data set with 500 cases is used to refit the Parsonnet score, the 95% confidence intervals for the estimates of \((\hat{\alpha}, \hat{\beta})^T\) are \((-4.36, -3.00)\) and \((0.025, 0.129)\), respectively, and, if there are 4,000 cases in the training data set, the respective 95% confidence intervals are \((-3.91, -3.45)\) and \((0.060, 0.094)\).

For both cardiac surgical populations, the approximate means, \(E(\hat{p})\), show a small downward bias with respect to \(\bar{p}\). The amount of the bias slowly decreases as the size, \(n\), of the training data set increases. The standard deviation, \(SD(\hat{p})\) of the recalibrated estimates \(\hat{p}\) for both the Parsonnet score and EuroSCORE are of the same order of magnitude as \(E(\hat{p})\) and it decreases marginally as \(n\) increases. The reason for this marginal decrease is that, although the value of \(E\{V(\hat{p} \mid X)\}\), as it is defined in Equation (6.35), halves as \(n\) doubles, it is at least an order of magnitude less than \(V\{E(\hat{p} \mid X)\}\). \(V\{E(\hat{p} \mid X)\}\) is relatively stable to changes in \(n\). It appears to slowly increase as \(E(\hat{p}) \to \bar{p}\).

### 6.5.2 Effect on ARL\(_0\)s of Risk Adjusted CUSUM Schemes

The second study on uncertainty in the estimation of model parameters investigated the effect that it has on the ARL\(_0\) performance measure. Approximate ARL\(_0\)s of the risk adjusted and minimum effect CUSUM schemes monitoring the mortality outcomes of the cardiac surgical populations given in Figures 5.3 (c) and 5.4 (c), where the respective patient populations are stratified according to the Parsonnet score and EuroSCORE, were found using simulation.

Uncertainty at the parameter estimation level was incorporated into the simulation process to approximate each ARL\(_0\) by making the following modifications of the process of simulating run lengths described in Section 5.3.3. The expected
probability of mortality, \( \hat{p}_t \), of patient \( t \), for \( t = 0, 1, \ldots \), was calculated by substituting the risk score, \( X_t \), and a parameter estimates, \( (\hat{\alpha}, \hat{\beta})^T \), drawn from the bivariate normal distribution described in Section 6.5.1. However, events 0 or 1 were drawn with \( 1 - p_t \) or \( p_t \), respectively, where \( p_t \) is computed using the “true” parameter values, \( (\alpha, \beta)^T \), as described in Section 6.5.1. Each ARL\(_0\) that was approximated with an uncertain parameter estimate \( (\hat{\alpha}, \hat{\beta})^T \) from 1,000 run lengths. The process was repeated with another \( (\hat{\alpha}, \hat{\beta})^T \) drawn from the bivariate normal distribution of interest until 500 approximate ARL\(_0\)s found.

Table 6.5 shows that there were four bivariate normal distributions used to model varying levels of uncertainty for the parameter estimates of the Parsonnet score and there were three for the EuroSCORE. In all, the modified simulation process was used to produce seven samples of 500 approximate ARL\(_0\)s for each of the risk adjusted and minimum effect CUSUM schemes. The results of some exploratory analysis are presented in the four charts in Figure 6.14 where the median and fifth and 95\(^{th}\) percentile values of the ARL\(_0\)s are plotted against the number of cases, \( n_P \) or \( n_E \), used in the training data sets described in Section 6.5.1. All median ARL\(_0\)s, except for the median in Chart (b) where the ARL\(_0\)s of the risk adjusted CUSUM scheme monitoring the EuroSCORE population were approximated from training data sets with 4,000 cases, are somewhat less than the approximations of the equivalent “true” ARL\(_0\)s given in Chapter 5. The fifth percentile of ARL\(_0\) values decreases and the 95\(^{th}\) percentile increases as the number of cases in the training data sets decreases. These results are consistent with the findings in Section 6.5.1 that there is a small downward bias in \( \hat{p} \) associated with uncertainty in parameter estimation and that \( V(\hat{p}) \) increases as the number of cases decreases.

The results and discussion in Sections 6.2–6.4 indicate that calibration of the expected probabilities, \( \hat{p}_t \), predicted by the risk model is an important influence on the ARL\(_0\) performance measure of risk adjusted and minimum effect CUSUM schemes, but the findings in this section indicate that the uncertainty in parameter estimates of the recalibration model is also a factor influencing the performance of both CUSUM schemes. The example used by Steiner et al. (2000) to illustrate the characteristics of the risk adjusted CUSUM scheme implies that the Parsonnet
Figure 6.14: Distributions of $\text{ARL}_0$s estimated with uncertain $(\hat{\alpha}, \hat{\beta})^T$ plotted against the size of training data sets used to estimate $(\hat{\alpha}, \hat{\beta})^T$. The asterisks give the median $\text{ARL}_0$ values and the filled and open triangles give the respective $5^{\text{th}}$ and $95^{\text{th}}$ percentiles of each $\text{ARL}_0$ distribution. Note the $\text{ARL}_0$s in Chart (b) and those in the other charts were approximated with the decision threshold $h = 3.5$ and 4.5, respectively.
score will accurately predict $\hat{p}_t$ after recalibrating it using Equation (6.29) and a training data set with approximately 2,000 cases, but this study showed that, after recalibration of Parsonnet score using 2,000 cases, the distribution of the approximate $\text{ARL}_0$s of the risk adjusted CUSUM scheme with $h = 4.5$ is such that $\Pr(2,551 \leq \text{ARL}_0 \leq 29,696) = 0.9$ because of the uncertainty in the estimation of the parameters, $(\hat{\alpha}, \hat{\beta})^T$ of Equation (6.29). If the number of cases in the training data set is doubled to 4,000, the distribution of the $\text{ARL}_0$s, approximated with uncertain $(\hat{\alpha}, \hat{\beta})^T$, remains flat with $\Pr(3,789 \leq \text{ARL}_0 \leq 20,240) = 0.9$.

Comparison of the distributions in Figures 6.14 (a) and (c) with their respective counterparts in Figures 6.14 (b) and (d) indicates that the $\text{ARL}_0$s of the minimum effect CUSUM schemes monitoring outcomes of the populations risk adjusted by the EuroSCORE and by the Parsonnet score are less sensitive to variability in estimates of $(\hat{\alpha}, \hat{\beta})^T$ for Equation (6.29) than the $\text{ARL}_0$s of risk adjusted schemes monitoring the mortality of both populations. If the distributions of the $\text{ARL}_0$s of both CUSUM schemes monitoring the outcomes of the EuroSCORE population, where $(\hat{\alpha}, \hat{\beta})^T$ was estimated using training datasets with 4,000 and 2,000 cases, are compared with the respective distributions of the $\text{ARL}_0$s of both CUSUM schemes monitoring the outcomes of the Parsonnet population, there is an indication that the $\text{ARL}_0$ performance indicator is more sensitive to the variability in $(\hat{\alpha}, \hat{\beta})^T$ when the EuroSCORE population was being monitored. It is possible that the reason for the increased sensitivity is that the mean probability of mortality, $\bar{p}$, of the EuroSCORE population is less than that of the Parsonnet score population.

### 6.5.3 Effect on Alarms from Risk Adjusted CUSUM schemes

The final study on the effect of variability of the parameter estimates, $(\hat{\alpha}, \hat{\beta})^T$, on the performance of risk adjusted CUSUM schemes uses the data for the upward and downward CUSUMs in Figure 5.2 to illustrate the effect of the variability of $(\hat{\alpha}, \hat{\beta})^T$ on the CUSUM itself. These data give the mortality outcomes, $Y_t \in \{0, 1\}$, and predicted probabilities of mortalities, $p_t \in (0, 1)$, of 2,398 patients admitted to an ICU between 1 January 1995 and 29 December 1997. Because neither the upward nor downward CUSUM chart signals between 11 November 1995 and 2 October
TABLE 6.6: Parameter estimates, $\hat{\alpha}$ and $\hat{\beta}$, for recalibration of the APACHE III. The estimates were derived by fitting training data sets of size $n$ to Equation (6.29).

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>150</td>
<td>-0.007</td>
<td>0.854</td>
</tr>
<tr>
<td>250</td>
<td>0.057</td>
<td>1.047</td>
</tr>
<tr>
<td>500</td>
<td>0.201</td>
<td>1.063</td>
</tr>
<tr>
<td>1,000</td>
<td>0.279</td>
<td>1.064</td>
</tr>
</tbody>
</table>

1996, it is assumed that the ICU process was in control during that period. The ICU data set was split into a training data set of 1,000 cases admitted to the ICU between 11 November 1995 and 2 October 1996 and a test data set of 1,398 cases admitted between 1 January and 10 November 1995 and between 3 October 1996 and 29 December 1997. An imaginary situation to explain the time of 11 November 1995 to 3 October 1996 for the collection of the training data in this study is that it was collected and used to recalibrate the APACHE III score when processes in the ICU were known to be in-control. The expected probabilities of mortality, predicted by the recalibrated risk model, are then used in a retrospective analysis of the mortality data collected during the early period and prospective monitoring of the mortality outcomes in the later period.

The logit transformation of the expected probabilities of mortality, $p_t$, gave a score, $X_t \in \mathbb{R}$, for each patient $t$ in both the training and test data sets. Parameter estimates $(\hat{\alpha}, \hat{\beta})^T$ for Equation (6.29), which were found by fitting it to $Y_t$ and $X_t$ in the training data set and subsets with 500, 250 and 150 cases, are given in Table 6.6. For the recalibrations with $n = 1,000$, 500, 250, and 150 cases, the respective predicted probabilities $\hat{\rho} < 0.013$, $\hat{\rho} < 0.040$, $\hat{\rho} < 0.228$ and $\hat{\rho} > 0.488$ were less than the equivalent expected probabilities of mortality given by the APACHE III risk score.

Upward and downward risk adjusted CUSUM charts, which use expected prob-
Figure 6.15: Examples to illustrate the influence of calibration of the APACHE III score on the path of upward and downward CUSUMs monitoring outcomes in an ICU. The benchmark paths in Plots (a) and (b) are computed using the uncalibrated score, the paths in Plots (c) and (d) are computed using scores recalibrated with training data with 250 patients, and the paths in Plots (e) and (f) are computed using scores recalibrated with training data with 500 patients.
Figure 6.16: Additional examples to illustrate the influence of calibration of the APACHE III score on the path of upward and downward CUSUMs monitoring outcomes in an ICU. The benchmark paths in Plots (a) and (b) are computed using the uncalibrated score, the paths in Plots (c) and (d) are computed using scores recalibrated with training data with 1,000 patients, and the paths in Plots (e) and (f) are computed using scores recalibrated with training data with 150 patients.
abilities of mortality calculated with APACHE III probabilities that were uncalibrated and recalibrated using the parameter estimates given in Table 6.6, are presented in Figures 6.15 and 6.16. The charts provide examples of the sensitivity of the risk adjusted CUSUM scheme to uncertainty in the expected probabilities of mortality predicted by the APACHE III score. The respective upward and downward CUSUM paths in Charts (a) and (b) of each figure, which were computed using the uncalibrated APACHE III score, are the benchmarks against which the paths calculated using the recalibrated scores may be compared.

The focus for the retrospective study is on the timing of the out-of-control signal for the upward CUSUM charts. For the benchmark CUSUMs charts, the alarm occurs at observation 583. The upward CUSUM in Figure 6.15 (c), where the risk adjusted CUSUM weights were calculated using expected probabilities were computed using the parameters \((0.057, 1.047)^T\) estimated from training data set of 250 patients, signals just before its benchmark at Observation 576. The CUSUMs in Figures 6.15 (e) and 6.16 (d), where the respective number of patients in the training data set was 500 and 1,000, signal after the benchmark at Observations 663 and 696, respectively. The most probable reason for the delayed alarms for an excessive number of deaths in the ICU is that the recalibrated expected probabilities are generally less than the probabilities predicted by the APACHE III score. Consequently, the CUSUM increments less than the benchmark if there is a death. This is well illustrated by the CUSUM chart in Figure 6.16 (e) where the probabilities of mortality were predicted after risk score recalibration using training data with 150 patients. Before the alarm from the benchmark CUSUM chart, patients 551, 560, 566, and 572, with respective expected probabilities of mortality predicted by APACHE III to be 0.011, 0.028, 0.018, and 0.027. If APACHE III is recalibrated using the data set with 150 patients, the predicted probabilities are almost double at 0.020, 0.046, 0.031, and 0.045. Therefore, increments are 0.682, 0.666, 0.675, and 0.667 for the benchmark CUSUM and 0.673, 0.648, 0.663, and 0.649 for the chart in Figure 6.16 (e). Consequently, there is no signal from the CUSUM in Figure 6.16 (e) because the magnitude of the positive CUSUM weights for these and other deaths is reduced and the magnitude of the negative weights for survival is increased.
All signals from the downward CUSUM charts occur after and all signals from the upward CUSUM charts occur prior to the collection of the test data used to recalibrate the APACHE III score. Thus, the prospective monitoring indicates that the proportion of deaths in the ICU is less than expected and provides evidence in support of an hypothesis of improved survival outcomes (Cook et al., 2002), but the timing of signals from the downward CUSUMs is sensitive to the calibration of the APACHE III score. The benchmark CUSUM chart plotted in Figures 6.15 (b) and 6.16 (b) signals at Observation 1,457, which is just after the signal at Observation 1,452 from the chart in Figure 6.15 (d), where the CUSUM weights calculated using expected probabilities from APACHE III recalibrated with training data with 250 patients. The weights for the other downward charts in Figures 6.15 (f), 6.16 (d), and 6.16 (f) were computed using probabilities computed from recalibrated APACHE III scores, where the number of patients in the training data set was 500, 1000, and 150, respectively. All these charts have multiple signals with each initial signal occurring before the signal from the benchmark chart at Observations 1,393, 1,347, and 1,354, respectively.

The upward and downward CUSUM charts presented in this illustrative study indicate the time and number of signals from risk adjusted CUSUM schemes show some sensitivity to the calibration of the risk score used to predict the expected probabilities of the adverse outcome being monitored. It is clear that the differences between the paths, that the benchmark CUSUM and CUSUMs for which increments were computed using recalibrated $\hat{p}_t$s follow, are related to the variability of each of the parameters $(\hat{\alpha}, \hat{\beta})^T$ estimated from one of the four training data sets to recalibrate the APACHE III score.

When Steiner et al. (2000) proposed the risk adjusted CUSUM scheme, they investigated its sensitivity to the uncertainty in estimating the parameters of Equation (6.29) by generating 1,000 bootstrap samples of 2,218 observations from their training data set by sampling with replacement. They used each sample to reestimate the parameters, $(\hat{\alpha}, \hat{\beta})^T$, of Equation (6.29). They then used the expected probabilities of mortality calculated using $(\hat{\alpha}, \hat{\beta})^T$ to plot the paths of risk adjusted CUSUM schemes monitoring the mortality outcomes following operations undertaken by two cardiac surgeons. Steiner et al. (2000) presented charts showing
the 5th and 95th percentiles of the generated CUSUM paths from which they concluded that the risk adjusted is relatively robust, although the 95th percentile of the CUSUM path did exceed the decision boundary on one occasion.

The CUSUM schemes in the example used by Steiner et al. (2000), which had an ARL₀ of 9,600, would not be expected to signal when monitoring the in-control outcomes for cardiac surgical operations undertaken by one surgeon over five years. Bridgewater et al. (2003) reported that the maximum number of operations undertaken by a single cardiac surgeon over a three year period was 598. Based on this report, the maximum number of observations in the CUSUM examples used by Steiner et al. (2000) is approximately 1,000. The results in Figure 6.14 show that, if a data set of 2,000 is used to estimate \((\hat{\alpha}, \hat{\beta})\)^T, the 5th percentile of the ARL₀ distribution for risk adjusted CUSUMs monitoring a population with \(\bar{p} = 0.046\) is given by ARL₀ \(\approx 2,560\). A risk adjusted CUSUM scheme with an actual ARL₀ of approximately 2,000 would not be expected to signal while monitoring 1,000 outcomes from an in-control process.

### 6.6 Conclusion

The results given in Tables 6.1 and 6.2 show that, if deaths are simulated with random probability \(p_t\), such that \(E(p_t) = p_m\), then the ARL₀s approximated for the minimum effect and risk adjusted CUSUMs were found to be equal to the ARL₀ estimated for deaths simulated with known probability \(p_m\). This finding suggests that, if the estimated probabilities of mortality, \(\hat{p}_t\), calculated using a risk model is well calibrated to the true probability of mortality \(p_t\), the approximations of the ARL₀ performance measure of any risk adjusted CUSUM schemes is unaffected by any uncertainty in the estimate \(\hat{p}_t\). However, the approximate ARL₀s found if deaths were simulated using the Gamma or Normal models described in Sections 6.2.2 and 6.2.4, respectively, indicated that the performance of both CUSUM schemes is quite sensitive to calibration of probabilities, \(\hat{p}_t\), predicted by the risk score, with respect to the “true” probabilities of mortality, \(p_t\). As the upward bias of \(\hat{p}_t\) with respect to \(p_t\) increased, the ARL₀s of the risk adjusted CUSUM schemes increased and became excessively large. If \(\hat{p}_t\) was biased downwards, the
ARL$_0$s of both CUSUM schemes decreased, approaching the ARL$_1$s where the CUSUM schemes would be unable to distinguish between out-of-control and in-control processes. It is clear that a well calibrated risk model is necessary for a satisfactory performance of the risk adjusted or minimum effect CUSUM scheme, but parameter estimation for recalibration of risk models is uncertain.

If a risk adjusted or minimum effect CUSUM scheme to monitor medical outcomes is introduced as part of a system to improve medical processes, its evaluation using a Phase I analysis of historical data and a Phase II monitoring, as recommended by Woodall (2006) in a discussion paper, may be considered. Other authors (Burkom, 2006; Steiner, 2006) also recommended use of Phase I/Phase II assessment of health care control charts. The Phase I analysis recommended by Montgomery (2001) for industrial production processes is to set trial control limits from an analysis of past data in conjunction with an audit of current procedures to ascertain if the processes are in-control or out-of-control. When reliable control limits have been established, the control chart is used to monitor production prospectively. This is sometimes called phase 2 of control chart usage. However, any initial set of control limits should always be regarded as trial limits and, in general, the effective use of a control chart requires periodic revision of its control limits. In the context of monitoring using risk adjusted and minimum effect control charts, retrospective analysis to recalibrate the risk score should be undertaken in conjunction with an audit of the medical processes being monitored and the expected probabilities of an adverse outcome should be reviewed regularly to ensure that the risk model remains well calibrated. However, the error in parameter estimation described in Section 6.5 complicates the decision that it is necessary to recalibrate the risk model. It is possible that an audit will reveal differences in risk model calibration that are due to random variation. In such circumstances, any recalibration of the risk model is unnecessary because the update is “chasing the noise”.

Further work is required to establish guidelines for effective application of risk models for monitoring outcomes in a quality program to improve medical processes and procedures. Issues arise because the collection of data in the quality setting is equivalent to collection of data for an observational study. In the simulation studies
in this chapter, two issues that were identified are the uncertainty in parameter estimation for recalibrating a risk score and biases introduced to the estimates of the expected probability of an adverse outcome by error in the measurement of the covariates of a logistic risk model.

6.7 Addendum

When the candidate investigated the effect of uncertainty in the estimation of the parameters of Equation (6.29) on the ARL₀ of risk adjusted CUSUM schemes, he was unaware of any publications about this research topic. However, one examiner advised that the literature on quality control in the industrial production context contains an analogous set of methods assessing the effect that estimation of linear regression parameters has on both ARL₀s and ARL₁s. This brief review is an introduction to the literature on the effects of parameter estimation on the performance of control charts.

A wide-ranging review of research on the effects of parameter estimation control chart properties was undertaken by Jensen et al. (2006). It provides summaries of publications investigating the effect of parameter estimation on the run length (RL) distributions of (i) Shewhart charts, including \( \bar{X} \) charts and charts (R and S) for dispersion, (ii) EWMA charts, (iii) CUSUM charts, (iv) chart for autocorrelated data, such as charts used to plot residuals of the data from one step ahead forecasts from an appropriate time-series model, for example, an AR(1) or ARIMA(1,0,1) model, (iv) Multivariate \( T^2 \) Control Charts (Hotelling, 1941), (v) Attribute Charts, and (vi) Charts for other situations, such as nonparametric charts that use a function of the largest order statistic as a control limit (Albers and Kallenberg, 2004). Summary values from the run length distribution, which allow effective evaluation of the performance of the charts include the ARL, the standard deviation of the run length (SDRL), and percentiles of the RL distribution.

Some conclusions in this review by Jensen et al. are:

- The assumption that the in-control values of the parameters are known simplifies the development and evaluation of control charts. In practice the
parameters are rarely known, and control charts are usually based on estimated parameters. When estimates are used in place of known parameters, the variability of the estimators can result in chart performance that differs from that of charts designed with known parameters.

- It is often impossible to tell a practitioner how a specific control chart constructed using estimated parameters will perform.

However, it is possible to construct hypothetical cases of RL distributions, conditional on the parameter estimates, to gain insight into the best and worst case performance scenarios for charts with estimated parameters.

- More data in Phase I are needed than is typically recommended to achieve performance comparable with the known parameters case.

- It is difficult to make an assessment of the impact of parameter estimation because it depends on the direction of the estimation error and the particular control chart setting. Jensen et al. state that this is a poor excuse for not making better comparisons of both conditional and marginal RL distribution performance, nor is it a justification for not using the SDRL and other percentiles of the RL distribution to supplement the ARL.

Cause-selecting control charts are examples of monitoring schemes that use linear regression. Wade and Woodall (1993) state that their purpose is to distinguish between incoming quality problems and issues in the current operations of industrial production processes and give the definitions and basic principles of cause-collecting control charts. Wade and Woodall give an example where the cause collecting scheme is based on finding a relationship between the incoming quality measures $X_t$ and the outgoing quality measures $Y_t$ from an industrial production process. As both $X_t$ and $Y_t$ are assumed to follow a normal distribution, a useful method of modelling the relationship is the simple linear regression model given by

$$Y_t = \alpha + \beta X_t + \epsilon_t,$$  \hspace{1cm} (6.36)

where $\epsilon_t$ is assumed independent with a $N(0, \sigma^2)$ distribution. Shewhart charts are used to monitor the incoming quality measure $X_t$ and the residuals $\epsilon_t$. If the
$X_t$ chart signals, the incoming process is out of statistical control, if the residual chart signals, the current operation is out-of-control, and, if both charts signal, both processes are out-of-control.

Shu et al. (2005) assessed the effect of the uncertainty in parameter estimation of the performance of such case-selecting schemes. They show that the in-control ARL distribution is skewed to the right and conclude that the optimal method of reducing the skewness is to decrease the uncertainty in the parameter estimates by increasing the size of the dataset used to fit the regression model. The candidate notes that the results of his study—on the effect of uncertainty in the estimation of the parameters of a simple logistic regression model on the ARL$_0$s of risk adjusted CUSUM schemes—given in Figure 6.14 also showed that the in-control ARL distribution is skewed to the right and that the amount of skewness is reduced as the size of the training dataset increases. In the event that the number of training data cannot be increased, Shu et al. recommend that monitoring commence using an appropriate self-starting scheme. Examples of self-starting CUSUM control charts are given in Hawkins and Olwell (1998, Chapter 7).
The material in this chapter is presented as it was when submitted to the journal, *Communications in Statistics—Simulation and Computation*. It has been published in that journal, see Webster and Pettitt (2007).

### 7.1 Introduction

A criterion for cumulative sum (CUSUM) schemes to monitor the number of non-conforming items in industrial manufacturing processes is the probability, $\pi$, that an item is non-conforming is constant (Hawkins and Olwell, 1998, Page 122). Such schemes are inappropriate for monitoring the number of adverse outcomes following medical procedures because each patient is unique and, clearly, the assumption, that each outcome following a medical procedure has constant probability, is false. The expected probability $\pi_t$ of an adverse outcome for patient $t$ may be estimated prior to the procedure using an appropriate risk model. For example, the Parsonnet score (Parsonnet et al., 1989) and the EuroSCORE (Nashef et al., 1999) are used to estimate the probability that a patient who undergoes a cardiac surgical operation
will die after the operation, where death is defined, for example, as in-hospital or within 30 days of the operation. In the context of monitoring medical outcomes, a risk model adjusts for variation in the patient population so that any alarm will be due to a change in the quality of treatment (Iezzoni, 1997). The risk-adjusted CUSUM (Steiner et al., 2000) is a scheme which is suited to monitoring adverse medical outcomes because it allows for patient variability.

The risk-adjusted CUSUM scheme signals an alarm when it crosses some predetermined decision boundary, $h$. It is possible for the CUSUM to signal when there is no shift in the outcome rate. This is a false alarm analogous to a false positive error in hypothesis testing. If there is a shift in the outcome rate, the time taken for a signal after the change occurs is analogous to the power of a hypothesis test. As in hypothesis testing, the performance of the CUSUM is a compromise between the time to false alarms and the time to true alarms (Hawkins and Olwell, 1998) so that the number of false alarms is tolerable but the response to an actual shift in the outcome rate is timely. Two useful but imperfect measures, imperfect because the run length distribution is highly variable (Hawkins and Olwell, 1998), of the performance of a CUSUM are the average run length (ARL) to an alarm when there has been no shift in the outcome rate and the ARL to an alarm after a shift in the outcome rate.

Grigg et al. (2003) note there are at least three ways to determine ARLs for CUSUM charts. Simulation is the most straightforward. It is time consuming and cumbersome but is useful when particular complexities of a chart, such as risk-adjustment and the discreteness of monitoring, make other approaches difficult. Another approach is to use numerical methods to solve an integral equation (Page, 1954) but, for more complex CUSUMs, complicated integral equations are difficult and, in some instances, impossible to solve. The final approach Grigg et al. (2003) describe is the Markov chain methodology used by Steiner et al. (2000) which provides a particularly convenient way to provide information on a variety of features of the run length distribution, such as the ARL, run length standard deviation, probability of crossing at or before a given time point, $t$, and higher moments. The Markov chain method requires all real values in the decision interval, $(0, h)$, to be discretized so, ideally, the number of Markov states of the transition probability
matrix should be as great as possible to minimize the error of the approximation. However, the degree of discretization is constrained by the computational intensity: the greater the mesh size of the transition matrix the greater the computer time required to manipulate it.

The method of computing the transition probability matrix described by Steiner et al. (2000) is close to that proposed by Brook and Evans (1972). In that method the continuum is approximated by placing the CUSUM $C_{t-1}$ at the centre of an interval $S_{t-1}$, concentrating all the probability at the centre of $S_{t-1}$. Fu et al. (2003) provide an example using this method, which we shall call rounding, and obtain accurate estimates of the ARL after 500 discretizations. On the other hand, Hawkins (1992) warns there is considerable experience that the rounding method of computing the transition probabilities leads to poor accuracy except at very fine discretizations. He proposes a more complex alternative of calculating the transition probabilities by smoothing the probability over the interval for the Markov state. This smoothing method is computationally attractive because it achieves accurate results using fewer discretizations than the rounding approach. Hawkins (1992) also warns that discrete jumps in the cumulative distribution function necessitate a finer mesh for an accurate final answer.

Our purpose is to show that calculation of the transition probability matrix using smoothing provides more accurate approximations of the ARLs of risk-adjusted CUSUMs using the Markov chain approach than those provided by calculation of transition probabilities using rounding. We only consider risk models where the predicted probabilities of adverse outcomes take a finite number of values. In Section 7.2 we review the risk-adjusted CUSUM scheme, introduce Hawkins’ method for computing the transition probabilities and adapt it for the risk adjusted CUSUM. In Section 7.3, we use some examples to show that, if transition probabilities are calculated using smoothing, ARL approximations using the Markov chain approach remain stable as the number of Markov states varies and converge to a limit as the number of discretizations increases but, if rounding is used, they are unstable and converge more slowly. In Section 7.4 we discuss the results.
## 7.2 Methods

The risk-adjusted CUSUM is used to monitor for a step increase, from \( p_0 \) to \( p_1 \), in the rate of adverse outcomes in a patient population. It takes the usual form

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

(7.1)

where \( C_t \) is the CUSUM at time \( t \), for \( t = 1, 2, \ldots \), and the CUSUM weight, \( W_t = \log \{l(y, p_1)/l(y, p_0)\} \), is the scoring found using the sequential-likelihood ratio test (Page, 1954). Observations, \( Y \), are assumed independent and there is an alarm if \( C_t \geq h \), where \([0, h)\) is the decision interval. It is possible to commence monitoring at any \( 0 \leq C_0 < h \) but we restrict our discussion to monitoring schemes that commence at \( C_0 = 0 \). In the context of medical procedure outcomes used for the discussion in this paper, the risk-adjusted CUSUM \( C_t \) relates to patient \( t \).

During the assessment prior to undergoing a medical procedure, the patient’s risk of an adverse outcome, \( X_t \) for patient \( t \), is scored by the medical practitioner who will undertake the procedure. A typical score is a finite, ordered scale of risk which takes integer values, \( x \) for \( x = 0, 1, \ldots, x_{\text{max}} \), where \( x_{\text{max}} \) is the largest value that the risk score takes. As shown in Section 7.2.1, the risk score may be used to estimate the expected probability \( p_t \) that patient \( t \) will experience an adverse outcome. When used to monitor the number of adverse outcomes, the risk-adjusted CUSUM allows for the varying \( p_t \) of the patient population by sequentially testing the hypotheses,

\[
H_0 : \frac{\pi_0/(1-\pi_0)}{p_t/(1-p_t)} = R_0,
\]

where \( \pi_0 \) is the probability of an adverse outcome for an in-control process and \( R_0 \) is the ratio of the odds of an adverse outcome for an in-control process to the expected odds of an adverse outcome after risk assessment of patient \( t \), versus

\[
H_1 : \frac{\pi_1/(1-\pi_1)}{p_t/(1-p_t)} = R_A,
\]

where \( \pi_1 \) is the probability of an adverse outcome for an out-of-control process and \( R_A \) is the ratio of the odds of an adverse outcome for an out-of-control process to the expected odds of an adverse outcome after risk assessment of patient \( t \).
From Steiner et al. (2000, Equation (2.3)), the weight \( W_t \) for observation \( Y_t \) of a risk-adjusted CUSUM scheme is

\[
W_t = \log \left[ \left\{ \frac{1 - (1 - R_0)p_t}{1 - (1 - R_A)p_t} \cdot \frac{R_A}{R_0} \right\}^{y_t} \left\{ \frac{1 - (1 - R_0)p_t}{1 - (1 - R_A)p_t} \right\}^{1-y_t} \right], \quad y_t \in \{0, 1\}.
\]  
(7.2)

### 7.2.1 Conditional Distribution of \( C_t \)

As Hawkins and Olwell (1998, Page 152) note, it follows from the assumed independence of the observations \( Y \) and the recursive definition of the CUSUM \( C_t \), given in Equation (7.1), that

\[
\Pr(C_t \mid C_0, C_1, \ldots, C_{t-1}) = \Pr(C_t \mid C_{t-1}).
\]

Thus the conditional probability of \( C_t \) is given by

\[
\Pr(C_t \mid C_{t-1}) = \Pr(W_t).
\]

From Equation (7.2), the event \( W_t = w \) is a function of the parameters \( R_0 \) and \( R_A \) and the random events that the expected probability of an adverse outcome \( p_t = p \) and the outcome \( Y_t = y_t \). Thus \( \Pr(W_t = w) \) is given by \( \Pr(Y_t = y_t, p_t = p) = \Pr(Y_t = y_t \mid p_t) \Pr(p_t = p) \). We assume that observed outcomes are distributed \( Y_t \sim \text{Bernoulli} \left( p_t \right) \) so that

\[
\Pr(Y_t \mid p_t) = p_t^{y_t} (1 - p_t)^{1-y_t}.
\]

Risks that patients will experience an adverse outcome are expressed on a linear scale, so a risk score \( X_t = x \) that patient \( t \) will experience an adverse outcome does not accurately reflect the probability of an adverse event. A reasonable model for the expected probability of mortality \( p_t \) of the \( t^{th} \) patient is given by

\[
\logit(p_t) = \alpha + \beta x,
\]  
(7.3)

where \((\alpha, \beta)^T\) is the regression parameter. For example, Steiner et al. (2000) used this model to calibrate the Parsonnet score for the expected probability of mortality for their example of the risk-adjusted CUSUM to monitor patient mortality.
following cardiac surgery. Now the patient population has some discrete distribution \( \Pr(X_t = x) \). Hence

\[
\Pr\{\text{logit}(p_t) = \alpha + \beta x\} = \Pr(X_t = x).
\]

Therefore the probability distribution of \( W_t \) is given by

\[
\Pr(W_t = w) = p_t^y (1 - p_t)^{1-y} \Pr(X_t = x),
\]

where the value of \( w \) for patient \( t \) is computed by substituting the value of \( p_t \), estimated by substituting the risk score \( x \) into Equation (7.3), and the value of \( y_t \) into Equation (7.2).

The Markov chain transition probabilities are found from the joint distribution of the observed mortalities, \( y_t \), and the risk scores, \( X_t \). We note that, for such a risk adjusted CUSUM, the weights are discrete, finite and take irrational values, \((w_1, \ldots, w_N)\), with a probability distribution \((v_1, \ldots, v_N)\) where \( \Pr(W_t = w_n) = v_n \) for \( n = 1, \ldots, N \). If the values of \( W \) are ordered so that

\[
w_1 < w_2 < \cdots < w_n < \cdots < w_N
\]

the cumulative probability distribution \( \Pr(W \leq w_n) = V_n \) is given by \( \sum_{j=1}^{n} v_j \).

### 7.2.2 Calculating Transition Probabilities Using Smoothing

When using the Markov chain property of CUSUMs to approximate their ARLs, the state space of the CUSUM, \( C_t \), is discretized into \( M + 1 \) states commencing at a reflecting state 0. State \( M \) is an absorbing state equivalent to \( C_t \geq h \). The width, \( \Delta \), of States 1 to \( M - 1 \), which are non-overlapping intervals in \((0, h)\), is \( \Delta = h/M \).

Brook and Evans (1972) show that the ARL may be found by solving the equation

\[
E(\lambda) = (I - R)^{-1}1,
\]

where \( E(\lambda) \) is the \( M \times 1 \) vector of expected run lengths to a signal, \( I \) is the \( M \times M \) identity matrix, \( R \) is the sub-matrix of the transition probability matrix excluding
transitions from or to the absorbing state, and \( \mathbf{1} \) is an \( M \times 1 \) vector with each element 1.

Steiner et al. (2000, Appendix) give a method of scaling and rounding off to calculate the transition probabilities of the matrix \( \mathbf{R} \) for risk-adjusted CUSUMs monitoring populations with discrete and finite categories of risk.

Smoothed transition probabilities may be computed using the equation

\[
\Pr(a < S_n < b \mid c < S_{n-1} < d) = \int_c^d \{F(b - s) - F(a - s)\} \, d\mu(s) \tag{7.5}
\]
given in Hawkins and Olwell (1998, Page 155). For a CUSUM moving from interval \( S_{t-1} = i \) to interval \( S_t = j \) we let \( \mu(x) \) be the distribution function of \( S_{t-1} \) conditional on \( (i - 1)\Delta < S_{t-1} < i\Delta \) and we have \( V(x) \) as the cumulative distribution of \( W \). Then Equation (7.5) becomes

\[
\Pr[(j - 1)\Delta < S_t < j\Delta \mid (i - 1)\Delta < S_{t-1} < i\Delta] = \int_{(i-1)\Delta}^{i\Delta} \{V[j\Delta - s] - V[(j - 1)\Delta - s]\} \, d\mu(s). \tag{7.6}
\]

Assume \( \mu \) to be uniform so that \( d\mu(s) = ds/\Delta \). For a transition from the \( i \)th to the \( j \)th Markov state we must have \( (i - 1)\Delta + w_n > (j - 1)\Delta \) or \( i\Delta + w_n < j\Delta \).

Suppose \( (i - 1)\Delta + w_n > (j - 1)\Delta \) then we define \( f \in (0, 1) \) such that

\[
f = \frac{w_n - (j - i)\Delta}{\Delta}.
\]

The cumulative distribution \( V(x) \) has a discontinuity at \( w_n = (j - i + f)\Delta \) where it steps by \( v_n \) from \( V_{n-1} \) to \( V_n \). Hence, Equation (7.6) may be evaluated as

\[
\Pr(S_t = j \mid S_{t-1} = i, W_n) = \int_{(i-1)\Delta}^{(i-1+f)\Delta} \{V[j\Delta - s] - V[(j - 1)\Delta - s]\} \, ds/\Delta \\
+ \int_{(i-1+f)\Delta}^{i\Delta} \{V[j\Delta - s] - V[(j - 1)\Delta - s]\} \, ds/\Delta \\
= (1 - f)v_n \tag{7.7}
\]

It is possible that, for a step of size \( w_n \), the transition is from \( S_{t-1} = i \) to \( S_t = j + 1 \). Then we find that Equation (7.6) evaluates as

\[
\Pr(S_t = j + 1 \mid S_{t-1} = i, W_n) = fv_n \tag{7.8}
\]
There are special cases. If the transition from the $i^{th}$ to the $j^{th}$ Markov state is such that $(i-1)\Delta + w_n > h$, then all $v_n$, the probability associated with the event $W_t = w_n$, accumulates with the transition probability $\Pr(S_t = M | S_{t-1} = i)$, where $S_M$ is the absorbing state, and if the transition is such that $i\Delta + w_n < 0$, then all $v_n$ accumulates with $\Pr(S_t = 0 | S_{t-1} = i)$, where $S_0$ is the reflecting boundary.

### 7.2.3 Simulation Method

The process of simulating run lengths of risk-adjusted CUSUMs schemes is by

- drawing a risk score at random from the population of scores;
- computing each patient’s probability, $p_t$, of an adverse outcome; for example, if the risk model is that the expected probability $p_t$ of an adverse outcome and the risk score $x_t$ have a logit relationship, $p_t$ is computed according to Equation (7.3);
- randomly generating outcome events $y_t \in \{0, 1\}$
  - for an in-control process, let $R_0 = 1$, then $\Pr(Y_t = 0) = 1 - p_t$ and $\Pr(Y_t = 1) = p_t$, or
  - for an out-of-control process, $\Pr(Y_t = 0) = (1 - p_t)/(1 + (R_A - 1)p_t)$ and $\Pr(Y_t = 1) = R_A p_t/(1 + (R_A - 1)p_t)$;
- monitoring the outcomes with the risk-adjusted CUSUM, with decision threshold $h$, tuned to signal if the odds ratio of observed to expected outcomes is $R_A$;
- recording the run length to a signal; and
- repeating until the prescribed number of run lengths recorded.

### 7.3 Results

An artificial study, Study 1, illustrates the method under an unrealistic risk score distribution. Suppose a risk-adjusted CUSUM scheme is used to monitor the
number of adverse outcomes following a medical procedure where the patient population may be stratified so that the risk of an adverse outcome has twenty-four discrete probabilities, 0.04, 0.08, . . . , 0.96, with uniform distribution. Approximate ARLs using the Markov chain approach were found where the decision threshold $h$ was set at 4.5 and the process out-of-control, and where $h$ was 3.5, 4.5 and 5.5 and the process in-control. Discretization of the decision interval $(0, h)$ increased by increments of 10 to a maximum of 3,000 Markov states.

The respective plots of the approximate ARL against the number of discretizations are shown in Rows A, B, C, and D of Figure 7.1. Column 1 gives ARLs
computed for the transition probabilities calculated using rounding and Column 2 gives ARLs for transition probabilities found using smoothing. The dashed or dotted lines on the respective plots in Columns 1 or 2, which are the lower and upper 95% confidence limits for the ARL of 100,000 run lengths simulated using the method described in Section 7.2.3, are used as a benchmark for the accuracy of the approximations found using the Markov chain approach. Simulations were done using the R statistical application (R Project, 2004) and estimates of ARLs using the Markov chain approach were computed with the MATLAB technical computing package (The Mathworks, 2004).

For the plots in Column 1, where rounding was used to compute transition probabilities, the ARL estimations vary erratically as the number of Markov states vary. In Plots A1 and B1 the ARLs estimated using the Markov chain approach are within the 95% confidence limits after 2,520 and 2,830 discretizations of \((0, h)\), but, in Plots C1 and D1, there are estimates outside the confidence limits as the number of Markov states approach 3,000. On the other hand, for Plots A2, B2, C2 and D2 where smoothing was used to calculate the transition probabilities, the estimates of the ARLs cross the lower 95% confidence bound for 150, 320, 570, and 940 Markov states, respectively, and remain within the confidence limits as the number of discretizations increase.

The degree of instability of the ARL approximations, which were calculated using rounded transition probabilities, appears to increase as the magnitude of the ARL increases. Table 7.1 shows that the maximum relative differences between the simulated ARL of each of the four CUSUM schemes in Study 1 and each of the equivalent ARLs approximated using the rounding approach increases from 1.12% to 9.68% as the “true” ARL increases.

The Parsonnet score (Parsonnet et al., 1989) and the EuroSCORE (Nashef et al., 1999) are two discrete risk scores used to predict the probability of mortality of patients undergoing cardiac surgical operations. In Study 2, we consider CUSUM schemes, risk-adjusted using these scores, to monitor the number of deaths following cardiac surgery. We approximate ARLs using the Markov chain approach and compare the estimates of the ARL found using rounding with those found using smoothing.
Table 7.1: Maximum relative differences between ARL approximations found using simulation and those found using the Markov chain approach with transition probabilities calculated using rounding.

<table>
<thead>
<tr>
<th>Distribution and Decision Threshold</th>
<th>Markov States†</th>
<th>ARL Estimate</th>
<th>Relative Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform*, $h = 4.5$</td>
<td>590</td>
<td>93.9</td>
<td>92.9</td>
</tr>
<tr>
<td>Uniform, $h = 3.5$</td>
<td>1,000</td>
<td>979.7</td>
<td>952.2</td>
</tr>
<tr>
<td>Uniform, $h = 4.5$</td>
<td>590</td>
<td>2,845.4</td>
<td>2,707.1</td>
</tr>
<tr>
<td>Uniform, $h = 5.5$</td>
<td>500</td>
<td>7,963.3</td>
<td>7,192.8</td>
</tr>
<tr>
<td>Parsonnet, $h = 4.5$</td>
<td>850</td>
<td>8,782.7</td>
<td>10,108.0</td>
</tr>
<tr>
<td>EuroSCORE, $h = 3.5$</td>
<td>510</td>
<td>8,112.1</td>
<td>11,781.0</td>
</tr>
</tbody>
</table>

†Number of discretizations ≥ 500 for maximum relative difference

*Out-of-control process

Figure 7.2: Cardiac surgical populations. Classified by (a) Parsonnet score, taken from Crayford (2000) and (b) EuroSCORE, taken from Bridgewater et al. (2003)
Figures 7.2(a) and (b) give the distributions of patient populations categorized by the Parsonnet score into 48 discrete levels of risk (Crayford, 2000) and the EuroSCORE into 20 levels of risk (Bridgewater et al., 2003), respectively. The probability of mortality for each patient category was found using Equation (7.3) where, for the Parsonnet score, the parameter \((\alpha, \beta)^T\) was taken as \((-3.68, 0.077)^T\) given by Steiner et al. (2000) and, for the EuroSCORE, as \((-5.56, 0.340)^T\) estimated using the data provided by Bridgewater et al. (2003) to fit a simple logistic regression model with the EuroSCORE as the explanatory variable and the mortality outcome as the response.

The plots in Figure 7.3 are as described for Figure 7.1 except that, for the plots in Row A, the monitoring scheme is a CUSUM risk-adjusted by the Parsonnet score and with the decision threshold \(h\) set at 4.5, and, for Row B, it is a CUSUM risk-adjusted by the EuroSCORE and with \(h\) at 3.5. In Column 1, where the transition probabilities are found using rounding, the ARLs of risk-adjusted CUSUMs vary unpredictably as the number of discretizations vary. Clearly, the instability of the ARLs of the CUSUM schemes risk-adjusted by cardiac surgical scores continues after 3,000 discretizations and it is more pronounced than any seen in Column 1 of Figure 7.1. Table 7.1 shows that the maximum relative differences between ARLs estimated using simulation and those approximated using the Markov chain approach are higher than any found for the CUSUM schemes in Study 1.

The values of the ARLs, approximated using the Markov chain approach with smoothing to calculate the transition probabilities, remain stable as the number of discretizations of \((0, h)\) vary. For risk-adjustment using the Parsonnet score, the approximate ARLs lie within the 95% confidence interval benchmark after 520 discretizations of \((0, h)\) (Figure 7.3 A2) and, for the EuroSCORE, after 450 discretizations (Figure 7.3 B2).

### 7.4 Discussion

In the two studies in Section 7.3, we compared two methods of calculating transition probabilities where the Markov chain approach is used to approximate ARLs of CUSUM schemes risk-adjusted using discrete, finite, risk models.
Figure 7.3: Study 2. The solid lines show the plot of approximate ARLs found using the Markov chain approach versus the number of discretizations of the decision interval \((0, h)\). The dotted lines give the upper and lower 95% confidence limits of the approximate ARLs found using simulation.
For transition probabilities calculated using rounding, we found the ARL approximations vary unpredictably as the number of divisions of the decision interval \((0, h)\) into discrete Markov states varies. The degree of instability decreases as the number of discretizations increases, but there are two other factors which may influence this instability. From Study 1, it is clear that there is more pronounced instability of the ARLs estimated using the Markov chain approach as the “true” value of the ARL of interest increases. A third factor is the distribution of the patient populations. The true values of ARLs for the CUSUM schemes risk adjusted by Parsonnet score and EuroSCORE and the scheme with uniform risk adjustment and \(h\) set to 5.5 are close to 8,000, but the distributions of the patient populations being monitored are different. The ARLs of the CUSUM risk-adjusted by EuroSCORE show greatest instability, and the ARLs of the scheme risk-adjusted by Parsonnet score show greater instability than the ARLs of the CUSUM with uniformly distributed risk-adjustment. We conclude there are at least three factors correlated with the degree of instability in the approximated ARLs; they are the number of discretizations of the decision interval, the magnitude of the ARL of the risk-adjusted CUSUM scheme, and distribution of the patient population categorized by risk score.

Where smoothing was used to compute the transition probabilities, the approximate ARLs of all the risk-adjusted CUSUM schemes studied converge smoothly from below to values that lie within the 95% confidence bounds for the ARLs estimated using simulation. For each scheme, less than 1,000 discretizations were required for the Markov chain approximation to be within the confidence interval benchmark.

Approximation of the ARLs of any CUSUM scheme using the Markov chain approach requires some numerical procedure to solve Equation (7.4). Relative confidence in the numerical solution depends on the conditioning of the matrix \((I-R)\) (Burden and Faires, 1997). We found that \((I-R)\) was ill-conditioned if either rounding or smoothing was used to compute transition probabilities. Despite the poor conditioning, MATLAB, which embeds “state of the art software for matrix computation” (MATLAB, 2004), provided stable solutions that were consistent with the ARLs found using simulation, if we calculated transition probabilities
using smoothing. We assume, therefore, MATLAB also provides numerically stable solutions to Equation (7.4) if rounding is used to compute transition probabilities and, consequently, it is the method of calculating transition probabilities which causes instability in the estimates of the ARLs.

Although this instability decreases as the number of Markov states increases, the accuracy of any one solution is uncertain. For example, for the CUSUM scheme risk-adjusted by EuroSCORE, we found that the ARL approximation with 2,580 discretizations of \((0, 3.5)\) exceeded the simulation estimate by 10.6%. On the other hand, the smoothing method outlined in Section 7.2.2 consistently provided good ARL approximations with less than 1,000 discretizations. It should be used to find transition probabilities if the Markov chain approach is used to estimate ARLs of CUSUM schemes risk-adjusted with discrete, finite risk-models.
Chapter 8

Summary and Conclusion

8.1 Results of Investigations

In the Overview in Chapter 1, it was stated that the purpose of the research undertaken for this thesis was to develop new statistical tools for surveillance and monitoring of adverse events in hospitals which adjust for differing patient and operational risk. The first part of this concluding chapter summarizes the results of the investigations into risk adjustment of mortality following cardiac surgery and the performance of the risk adjusted CUSUM scheme that were undertaken to achieve that goal.

8.1.1 Cardiac Surgical Risk Scores

There is an extensive literature on risk adjustment of all cardiac surgical outcomes. Chapter 3 provides a comprehensive literature review of cardiac surgical risk scores. It was noted that the discriminating ability of a risk model appeared to be independent of the number of risk factors used as explanatory variables. It was proposed that each of the risk factors in the cardiac surgical risk models for mortality be classified as one of the seven meta-factors categories or the mis-
cellaneous factor which are defined in Table 3.3. In fact, it was proposed that every risk factor classified as a meta-factor is a measure for one of the dimensions of risk—physiological reserve, cardiac reserve, diagnosis, and acute illness—given in Table 3.3. The meta-factor classes and the associated dimensions of risk were proposed as “medically meaningful” and were assumed to provide independent measures for the risk of cardiac surgical mortality. Clearly, further work with cardiac surgical experts and statistical investigations of risk adjustment models for cardiac surgical mortality will be needed to support or disprove those assumptions.

The cardiac surgical risk scores reviewed were developed from a single database, but comparative studies of the performance of the Parsonnet score and EuroSCORE suggest that risk scores should not be generalized to patient populations remote from their developmental databases if accurate predictions of the probability of mortality are required (Gogbashian et al., 2004). There are many factors—some of which can be measured and some which cannot—which affect the expected probabilities of mortality (Lilford et al., 2004). They identify the definitions and quality of data as issues. For example, in the literature review of cardiac surgical scores undertaken here, there are three different definitions of postoperative mortality. Each definition did provide a suitable measure of short term mortality for use as part of a monitoring process to assess the quality of surgical care. However, the differences in the definition of mortality would result in different counts of deaths following surgery and, therefore, the likelihood that an uncalibrated risk score would over- or under-predict a patient’s expected probability of mortality.

The conclusion was that institutions could use an “off the shelf” risk adjustment score to obtain accurate estimates of the probabilities of mortality of their cardiac surgical patients by using data collected from their patients to recalibrate the risk model. Such data should be collected in conjunction with a properly implemented review undertaken by the surgical teams and other interested parties to ensure that the recalibrated risk score provides expected probabilities of mortality for in-control surgical processes. Alternatively, the notion of meta-factors could be developed so that institutions could use their data to construct customized risk scores with risk factors chosen from variables routinely collected within the institution. The magnitude of the measurement error and problems of definition of risk factors for
such customized models are likely to be reduced. That is, there is a greater chance that data used for predicting mortality are consistent and reliable.

8.1.2 Multi-Institutional Monitoring Schemes

Currently, health care administrations, such as the New York State Health Department, publish information on adverse outcomes in the form of report cards. According to the New York State Health Department, the purpose of publishing RAMRs for cardiac surgical patients is that it can be used to enhance the quality of care. However, the publication of RAMRs in the form of report cards is criticized because there is a marked variation in the rankings of hospitals from year to year and medical practitioners are generally sceptical about the quality of data used to calculate the RAMRs. This uncertainty associated with RAMR estimation may be mitigated by the use of sequential analysis because such analysis, which uses data collected over time, provides more information about the performance of each institution. On the other hand, report cards provide a cross-sectional summary of the performances of all institutions for one time period only.

The medical practitioners, who are the subject of the report cards, have objected to their publication for many reasons. Some objections, such as the delays in their publication mean users of the report can have little confidence that the ratings are still applicable, appear well supported by the evidence. Others, such as publication of outcomes lead to outsourcing of high risk patients, seem less well supported by the evidence. Despite the objections from medical practitioners, report cards continue to be published as part, it is stated, of a quality improvement process. It is clear, however, that another reason for publication of report cards by health care administrations is to have a process of public accountability for the quality of care that hospitals deliver. In such circumstances, it is expected that publication of such indicators of the “quality” of health care will continue. Therefore, it is important that medical organizations and other professional bodies become involved in the development of those quality indicators to ensure they are meaningful and used in a positive manner, as the Royal College of Physicians did for the publication of report cards on the treatment of heart attack patients in
8.1.3 Performance of Risk Adjusted CUSUM Schemes where Outcome Rates are Assumed Known

The average run lengths, \( \text{ARL}_0 \)s, where the process is in-control, and \( \text{ARL}_1 \)s, where the process is out-of-control, were chosen as the measures for assessing risk adjusted CUSUM schemes. In the first study, ARLs were approximated using a simulation process in which non-adverse events 0 and adverse events 1 were generated for patient \( t \) with probability \( 1 - p_t \) and \( p_t \), respectively, where \( p_t \) was the probability predicted by the risk score that patient \( t \) would experience an adverse event.

The results of the simulation studies showed that both the \( \text{ARL}_0 \)s and \( \text{ARL}_1 \)s of risk adjusted CUSUM schemes increased markedly as the average expected probability of mortality of the patient population, \( \bar{p} \), for \( \bar{p} < 0.05 \), decreased (see Figures 5.6 and 5.7). The candidate hypothesized that a decrease in \( \bar{p} \) causes increases in the values of both ARLs. This hypothesis is supported by the statistical theory that the average sample number (ASN) of a sequential probability ratio test (SPRT) increases as the treatment effect size decreases (Jennison and Turnbull, 2006). The risk adjusted CUSUM scheme may be regarded as a sequence of SPRTs set to detect a step from \( p \) to approximately \( R_A p \), for small \( p \). As the CUSUM step size \( |R_A p - p| \) decreases, both \( \text{ARL}_0 \) and \( \text{ARL}_1 \) increase as \( p \) decreases. The candidate designed a risk adjusted CUSUM scheme with the step size constrained to be no less than a minimum value of 0.05 for all \( p < 0.97 \). Both its \( \text{ARL}_0 \) and \( \text{ARL}_1 \) were found to be stable under varying \( \bar{p} \). The minimum effect CUSUM scheme is recommended for monitoring the outcomes of medical treatments and procedures where the average adverse outcome rate is less than 0.05.

8.1.4 Performance of Risk Adjusted CUSUM Schemes where Outcome Rates are Uncertain

Simulation studies were undertaken to investigate the effect of uncertainty in the predicted probabilities of mortality at the individual level and at the model level.
on the ARL\(_0\)s of risk adjusted and minimum effect CUSUM schemes.

Uncertainty at the individual level was manufactured by drawing the “true” probability of mortality, \(p_T\), or \(g(p_T)\), where \(g(\cdot)\) is some function, from an appropriate probability distribution and generating the events 0 or 1 from \(p_T\). Provided the values of \(p_T\) were unbiased with respect to the predicted probabilities derived from the risk model, \(p_m\), the approximate ARL\(_0\)s found with deaths simulated with probability \(p_T\) were no different from the approximate ARL\(_0\)s found using the events 0 or 1 generated from \(p_m\), where \(p_m\) is assumed to be the known true probability of patient mortality. However, the performance of both CUSUM schemes was quite sensitive to any bias in \(p_T\). Thus, the key finding from this study of patient-level variability in expected probabilities predicted by risk adjustment scores is that calibration of the risk score to the population being monitored is an important issue for the performance of adjusted CUSUM schemes. If the calibration is poor, it is likely that either the number of false alarms will be excessive or the CUSUM will not give timely signals when the process is truly out-of-control.

A second study investigated the effect of uncertainty in the calibration of the Parsonnet score and the EuroSCORE on the ARL\(_0\)s of risk adjusted and minimum effect CUSUM schemes. Uncertainty was introduced at the model level by assuming that the parameter estimates \((\hat{\alpha}, \hat{\beta})^T\) found for the recalibration equation

\[
\logit(p_t) = \alpha + \beta X_t,
\]

where \(p_t\) is the expected probability of mortality and \(X_t\) is the risk score of patient \(t\), were from a bivariate normal distribution with mean \((\alpha, \beta)^T\) and variance dependent on the number of training cases used to fit the calibration equation. The ARL\(_0\)s of both the risk adjusted and minimum effect CUSUM schemes were approximated using a simulation process in which the events 0 and 1 were generated with probability \(p_t\) derived from the calibration equation with parameters \((\alpha, \beta)^T\), but the CUSUM weights were computed with expected probability of mortality \(\hat{p}_t\) derived from the calibration equation with parameter estimates \((\hat{\alpha}, \hat{\beta})^T\). The process was repeated 500 times to obtain sample distributions of the approximate ARL\(_0\)s for uncertain \((\hat{\alpha}, \hat{\beta})^T\).

The results showed that the range between the 5\(^{th}\) and 95\(^{th}\) percentiles of the
ARL\textsubscript{0} distributions decreased as the number of patients for the training data to recalibrate the risk scores increased. However, there was still considerable variation in the approximated ARL\textsubscript{0}s found using parameters (\(\hat{\alpha}, \hat{\beta}\))^T estimated using the highest number of cases in the training data. The variability in the parameter estimates found in this study indicates there is a need to monitor observed and predicted outcomes to ensure that risk models used in schemes for the surveillance of adverse medical outcomes are recalibrated if there is evidence that the current model calibration is inadequate.

### 8.1.5 Approximating Stable ARLs of Risk Adjusted CUSUM Schemes Using the Markov Approach

In this thesis, the studies to approximate the ARLs of risk adjusted CUSUM schemes were undertaken using the simulation approach because the Markov approach proposed by Steiner et al. (2000) of discretizing the decision interval and approximating the continuum by placing the CUSUM at the centre of a discretized interval for a Markov state was found to be unstable. An alternative method, which calculated transition probabilities by smoothing the probability over the interval for the Markov state, achieved more stable results with fewer discretizations, but it is limited to risk adjusted CUSUM schemes where the risk model is finite and discrete. This chapter was recently published as a paper in *Communications in Statistics: Simulation and Computation* (Webster and Pettitt, 2007).

### 8.2 Possible Further Research

#### 8.2.1 Cardiac Surgical Risk Models

The classification of the risk factors used for prediction of patient mortality following cardiac surgery into their meta-factors is a prototype which was proposed in Chapter 3 after consultation with an intensive care specialist with expertise in risk models for the mortality of patients admitted to intensive care. The next step is to present this concept of categorizing cardiac surgical risk factors into meta-factors and the classification used for this research to cardiac surgeons and clinicians.
Their opinions should be sought as to the appropriateness of and modification to the current classification.

The results in Chapter 6.3 showed that, if variation in logit($p_T$) is unbiased with respect to logit($p_m$), where $p_T$ is the true expected probability of mortality and $p_m$ is the expected probability predicted by the risk model, $p_T$ is biased upwards with respect to $p_m$. For risk models based on logistic regression, a possible source of unbiased variation in the logit dimension is measurement error of the covariate risk factors. Investigations to quantify any measurement errors of risk factors in logistic models and their effects on the probabilities predicted by the risk models could be undertaken. Expected probabilities of mortality may be derived using alternative methodologies, such as artificial neural networks and support vector machines (Cook et al., 2004), a classification tree approach (Graham et al., 2007), and propensity score risk adjustment (Stukel et al., 2007). Investigation of the effect of error in the measurement of explanatory variables on the estimated probability of mortality could also be undertaken.

There is some preliminary calculation of the effect of uncertainty in the estimates of the parameters ($\alpha, \beta$)$^T$ of the recalibration equation given Equation (6.29) on the mean $\bar{p}$ and variance Var($\bar{p}$) of the predicted probabilities of mortality. The component of Var($\bar{p}$) derived from the uncertainty of the parameter estimates, given in Equation (6.35) as $E\{V(\hat{\bar{p}}|X)\}$ is relatively small compared with the component derived from the risk mix of the patient population, given as $V\{E(\hat{\bar{p}}|X)\}$. The small effect on Var($\bar{p}$) from the uncertainty in ($\hat{\alpha}, \hat{\beta}$)$^T$ is difficult to reconcile with their large effect on the ARLs of the risk adjusted and minimum effect CUSUM schemes. Further research is needed to confirm the results given in Table 6.5.

8.2.2 Other than Patients Characteristics’ Measures of Risk

Lilford et al. (2004) identified definitions of variables, data quality, case-mix, clinical quality of care, and chance as classes of factors that affect the variability in the expected probability of the outcome of interest. They state that each of these factors have components that can be measured and those that cannot, but
case-mix (risk adjustment) is by far the most widely method to analyse the variability in outcome. In fact, in the literature reviews undertaken for the research in this thesis, the risk models only adjusted for the risk of adverse outcomes by using measures of the patients’ characteristics for increased risk of the outcome in question.

Randomized controlled trials are considered the gold standard for the evaluation of questions in medicine, but they may be inappropriate when trying to accurately quantify the frequency of rare events, such as mortality following a cardiac surgery (Wunsch et al., 2006). They state that observational studies can provide invaluable information, provided they are properly constructed with a clear understanding of the role of bias and confounding in the data and its interpretation. An appropriate study design deals with biases by adjusting for the confounders, provided information on the confounding variable is available and can be taken into account. Otherwise, the problem of bias from the unmeasured confounder is insurmountable. Wunsch et al. advise sensitivity analysis may be used to estimate how large an effect that an unknown confounder would have to have to eliminate or produce the effect seen.

One source of bias that may be measured is the differences in the definitions of the variables used for the risk models. For example, it is noted in Section 8.1.1 that there were three definitions of death in the review of risk adjustment for mortality following cardiac surgery. Clearly, this source of bias could be markedly reduced if all interested workers in the health care industry agreed to a convention for defining the response variables and risk factors for adverse outcomes. Another source of bias is that the observed mortality rate following cardiac surgery decreases over time. In Chapter 4.3.2, learning curve improvement in surgical techniques was suggested as a possible reason for the decrease in postoperative mortality. If this hypothesis is correct, it should be possible to include a variable or variables to adjust for this bias. Aylin et al. (2007) used the year of operation as a covariate for decreased risk of mortality following medical procedures, but the relationship between decreased adverse outcomes and improved technique may be more complex because of the non-linear nature of the learning curve.

Wunsch et al. (2006) give five techniques—matching, stratification, multivari-
able adjustment, propensity scores, and instrumental variables—as epidemiological tools to address the problem of confounding in epidemiological data. They caution that these techniques are primarily useful for dealing with measured, or known confounders. In this thesis, the severity of illness, or risk score, is the class of multivariable adjustment model which was considered in detail. Wunsch et al. state propensity scores are similar to severity of illness scores as they provide a single number to represent a large number of variables. The difference is that a propensity score is constructed to represent the probability of a patient encountering the exposure of interest by taking into account many variables. Propensity scores may be used to adjust for confounding factors by matching individuals with the same scores or by stratifying patients by the score and calculating the “pooled stratum specific estimates”. Wunsch et al. note that propensity scores have the advantage of modelling what actually happens. However, they state that, when propensity scores and multivariable regression were compared, there was no statistical difference between the effect estimates in 90% of cases.

The final technique, which Wunsch et al. (2006) propose as a means to account for bias and confounding, is instrumental variables. Wunsch et al. note that this technique is rarely used by the medical community. It is typically applied by health economists to examine quality of care. For example, they suggest that instrumental variables might be used for estimating the differences in quality of care between hospitals after accounting for the possibility that patients who are sicker tend to choose a better hospital. According to Wunsch et al., the most common problem with the use of instrumental variables is finding a variable that everyone agrees does not correlate with the outcome of interest.

Stukel et al. (2007) compared four analytical methods—multivariable model risk adjustment, propensity score risk adjustment, propensity based matching, and instrumental variable analysis—for removing the effects of selection bias in observational studies. They used a national cohort of elderly patients in the United States to show that the measure of the association between long term survival following AMI and cardiac catheterization was sensitive to the analytic method used. They concluded that instrumental variable analysis may produce less biased estimates of treatment effects, but is more suited to answering policy questions
than specific clinical questions.

It has been suggested in this thesis that governance is an important reason for publication of report cards comparing health care institutions. Therefore, econometric analytical tools may be appropriate for removing bias from the observational data used to compare institutions. Stochastic frontier analysis (Kumbhakar and Knox Lovell, 2000) is one such tool that may be used to develop measures of comparative performance in order to set a level playing field for health care providers (Jacobs et al., 2006). In health care, analysis using this statistical tool would be undertaken to assess the providers’ use of available resources to optimize various outputs (Kumbhakar and Knox Lovell, 2000). Stochastic frontier analysis may be a means of providing high-quality information in public domain in order to enhance health care providers’ accountability for the services they offer (Jacobs et al., 2006).

8.2.3 Monitoring Schemes

Cook et al. (2003) showed that risk adjusted CUSUM schemes could be used to monitor deaths in an intensive care unit where the patient mortality rate is usually greater than 10%. In many medical procedures, however, the death rate is approximately 0.01 or even less. This thesis has shown that, in such circumstances, doubling of the mortality rate is difficult to detect if the risk adjusted CUSUM is used for the surveillance. It may be possible to design risk adjusted schemes, based on the geometric or negative binomial distributions, that plot the number of operations, say, until there are $k$ adverse events, but the number of operations undertaken at an institution or by a surgeon may be insufficient for such charts to be effective.

It is possible that control charts to monitor the number of mortalities following surgery, such as cardiac surgery, are ineffective because the mortality rate is so low. An alternative surveillance scheme is to monitor quality characteristics for key clinical processes as the Royal College of Physicians do for treatment of patients diagnosed with acute myocardial infarction who are admitted to hospitals in England and Wales (see Chapter 4.4). In the case of cardiac surgery, it would
be expected such a monitoring scheme would be implemented with the support of the appropriate professional body and, because the mortality rate for cardiac surgery is very low, any deaths would be independently audited. A possible audit process is the Scottish Audit of Surgical Mortality described by Thompson and Stonebridge (2005).

8.3 Closing Comments

The research presented in this thesis has focused on monitoring the number of mortalities following cardiac surgery using the risk adjusted CUSUM chart, where the purpose is to ensure timely warning of surgical process going out-of-control. In Section 8.2.3, further work to develop a monitoring regime in which each death would be independently audited was proposed. In that case, controls charts to monitor clinical processes might also be implemented. However, it has become clear that monitoring for quality improvement can only be successfully implemented in conjunction with a quality improvement program. For example, much of the research in this thesis concerned monitoring the outcomes of cardiac surgery. Clearly, implementation of any monitoring scheme as part of quality improvement programs for cardiac surgery requires collaboration with cardiac surgeons and their professional organizations, such as The Australasian Society for Cardiac and Thoracic Surgeons (ASCTS, 2006).

The health departments in Australian states have also established units for improving the quality of health care services. For example, New South Wales Health has established The Greater Metropolitan Clinical Task Force (GMCT, 2004) with the objective of improving health care in hospitals by having specialty based clinical networks identifying how and where improvement might be made for their specialities and Queensland Health has the Quality Measurement and Strategy Unit (DRAC, 2007) which produces the Queensland Health report card described in Chapter 4. Any quality improvement program in Queensland Public Hospitals would have the greatest chance of success if there were agreement and liaison between the Queensland Health hierarchy to which the Quality Measurement and Strategy Unit belongs and the relevant professional organizations.
# Appendix A

## Risk Factors and Associated Meta-Factors

The 181 risk factors found in 31 cardiac surgical risk models (Djamaludin, 2006) and their meta-factors classes as given in Table 3.3 are listed in this appendix.

<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. &gt; 90% narrowing of left main trunk</td>
<td>3</td>
</tr>
<tr>
<td>2. 1-sec expiratory rate (% of normal)</td>
<td>2</td>
</tr>
<tr>
<td>3. Abdominal Aortic Aneurysm</td>
<td>2</td>
</tr>
<tr>
<td>4. Active AIDS</td>
<td>2</td>
</tr>
<tr>
<td>5. Active cancer</td>
<td>2</td>
</tr>
<tr>
<td>6. Active Endocarditis</td>
<td>7</td>
</tr>
<tr>
<td>7. Acute aortic dissection</td>
<td>6</td>
</tr>
<tr>
<td>8. Acute preoperative state (in cardiogenic shock, or intra-aortic balloon pump inserted preoperatively not for prophylactic purposes)</td>
<td>7</td>
</tr>
<tr>
<td>9. Age</td>
<td>1</td>
</tr>
<tr>
<td>10. Age of myocardial infarction</td>
<td>3</td>
</tr>
<tr>
<td>11. Albumin &lt; 4.0mg/dl</td>
<td>2</td>
</tr>
<tr>
<td>12. Anemia (Hematocrit ≤ 0.34)</td>
<td>7</td>
</tr>
</tbody>
</table>

*continued next page*
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>13. Angioplasty failure</td>
<td>7</td>
</tr>
<tr>
<td>14. Aortic dissection</td>
<td>6</td>
</tr>
<tr>
<td>15. Aortic surgery</td>
<td>6</td>
</tr>
<tr>
<td>16. Aortic surgery and Pressure Gradient</td>
<td>6</td>
</tr>
<tr>
<td>17. Atrial arrhythmia</td>
<td>5</td>
</tr>
<tr>
<td>18. AVR</td>
<td>6</td>
</tr>
<tr>
<td>19. Body mass index $&lt; 24$ kg.m$^{-2}$</td>
<td>2</td>
</tr>
<tr>
<td>20. Body surface area (m$^2$)</td>
<td>2</td>
</tr>
<tr>
<td>21. CABG at the time of valve surgery</td>
<td>6</td>
</tr>
<tr>
<td>22. Cachexia</td>
<td>2</td>
</tr>
<tr>
<td>23. Canadian cardiovascular society class 1-4 angina</td>
<td>3</td>
</tr>
<tr>
<td>24. Cardiac insufficiency</td>
<td>4</td>
</tr>
<tr>
<td>25. Cardiogenic shock</td>
<td>4</td>
</tr>
<tr>
<td>26. Cardiomegaly (x-ray)*</td>
<td>4</td>
</tr>
<tr>
<td>27. Cardiopulmonary resuscitation</td>
<td>7</td>
</tr>
<tr>
<td>28. Carotid Arterial disease</td>
<td>2</td>
</tr>
<tr>
<td>29. Catastrophic states (eg. Acute structural defect, cardiogenic shock, acute renal failure)</td>
<td>7</td>
</tr>
<tr>
<td>30. Catheterization-induced coronary closure</td>
<td>7</td>
</tr>
<tr>
<td>31. Cerebral vascular disease</td>
<td>2</td>
</tr>
<tr>
<td>32. Cerebrovascular accident</td>
<td>2</td>
</tr>
<tr>
<td>33. Cerebrovascular disease</td>
<td>2</td>
</tr>
<tr>
<td>34. Charlson Comorbidity Score</td>
<td>2</td>
</tr>
<tr>
<td>35. CHF</td>
<td>4</td>
</tr>
<tr>
<td>36. Chronic obstructive lung disease</td>
<td>2</td>
</tr>
<tr>
<td>37. Chronic pericarditis</td>
<td>5</td>
</tr>
<tr>
<td>38. Chronic pulmonary obstructive disease</td>
<td>2</td>
</tr>
<tr>
<td>39. (Chronic) renal failure</td>
<td>2</td>
</tr>
<tr>
<td>40. Chronic renal insufficiency (creatinine $&gt; 1.9$ mg/dl)</td>
<td>2</td>
</tr>
<tr>
<td>41. Combined CABG and mitral valve surgery</td>
<td>6</td>
</tr>
<tr>
<td>Risk Factor</td>
<td>MF</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>----</td>
</tr>
<tr>
<td>42. COPD</td>
<td>2</td>
</tr>
<tr>
<td>43. Coronary dissection</td>
<td>7</td>
</tr>
<tr>
<td>44. Critical preoperative state</td>
<td>7</td>
</tr>
<tr>
<td>45. Current digoxin use</td>
<td>4</td>
</tr>
<tr>
<td>46. Current diuretic use</td>
<td>4</td>
</tr>
<tr>
<td>47. Date of operation</td>
<td>8</td>
</tr>
<tr>
<td>48. Diabetes (unspecified type)</td>
<td>2</td>
</tr>
<tr>
<td>49. Diabetes mellitus (insulin-dependent/nondependent)</td>
<td>2</td>
</tr>
<tr>
<td>50. Diabetic status</td>
<td>2</td>
</tr>
<tr>
<td>51. Dialysis dependent (Peritoneal Dialysis or Hemodialysis)</td>
<td>2</td>
</tr>
<tr>
<td>52. Diffusely narrowed coronary arteries</td>
<td>3</td>
</tr>
<tr>
<td>53. Disaster</td>
<td>8</td>
</tr>
<tr>
<td>54. Ejection fraction</td>
<td>4</td>
</tr>
<tr>
<td>55. Electrocardiographically determined left ventricular hypertrophy</td>
<td>5</td>
</tr>
<tr>
<td>56. Elevated blood urea nitrogen (&gt; 29mg/dl)</td>
<td>2</td>
</tr>
<tr>
<td>57. Emergency surgery</td>
<td>6</td>
</tr>
<tr>
<td>58. Emergency surgery following PTCA/ catheterization complications</td>
<td>6</td>
</tr>
<tr>
<td>59. Emergent/salvage</td>
<td>6</td>
</tr>
<tr>
<td>60. Evident/Congestive heart failure</td>
<td>4</td>
</tr>
<tr>
<td>61. Evolving myocardial infarction</td>
<td>7</td>
</tr>
<tr>
<td>62. Extent of coronary disease</td>
<td>3</td>
</tr>
<tr>
<td>63. Extracardiac arteriopathy</td>
<td>2</td>
</tr>
<tr>
<td>64. HDL levels</td>
<td>2</td>
</tr>
<tr>
<td>65. Hemodynamic instability</td>
<td>7</td>
</tr>
<tr>
<td>66. Hypertension (systolic Blood Pressure &gt; 140mmHg)</td>
<td>2</td>
</tr>
<tr>
<td>67. Hyponatremia</td>
<td>7</td>
</tr>
<tr>
<td>68. Hypotension and shock</td>
<td>7</td>
</tr>
<tr>
<td>69. Idiopathic thrombogenic purpura</td>
<td>2</td>
</tr>
</tbody>
</table>

*continued next page*
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>70. Immunosuppressive medication</td>
<td>2</td>
</tr>
<tr>
<td>71. Incomplete revascularization</td>
<td>6</td>
</tr>
<tr>
<td>72. Inotropic support</td>
<td>7</td>
</tr>
<tr>
<td>73. Insulin-treated diabetes</td>
<td>2</td>
</tr>
<tr>
<td>74. Intra-aortic balloon pulsation</td>
<td>6</td>
</tr>
<tr>
<td>75. Intravenous nitrate</td>
<td>6</td>
</tr>
<tr>
<td>76. Intravenously administered inotrope</td>
<td>6</td>
</tr>
<tr>
<td>77. Intravenously administered nitroglycerin</td>
<td>6</td>
</tr>
<tr>
<td>78. Ischemic mitral incompetence</td>
<td>6</td>
</tr>
<tr>
<td>79. Jehovah’s Witness</td>
<td>6</td>
</tr>
<tr>
<td>80. Left Main CAD</td>
<td>3</td>
</tr>
<tr>
<td>81. Left main coronary artery involvement (stenosis)</td>
<td>3</td>
</tr>
<tr>
<td>82. Left main disease $&gt; 90%$</td>
<td>3</td>
</tr>
<tr>
<td>83. Left Ventricular Dysfunction</td>
<td>4</td>
</tr>
<tr>
<td>84. Left ventricular aneurysm (resected)</td>
<td>5</td>
</tr>
<tr>
<td>85. Left ventricular ejection fraction (Grade 1, 2, 3 or 4)</td>
<td>4</td>
</tr>
<tr>
<td>86. Left ventricular end-diastolic pressure</td>
<td>4</td>
</tr>
<tr>
<td>87. Left ventricular hypertrophy</td>
<td>5</td>
</tr>
<tr>
<td>88. Long term corticosteroids or immunosuppressive therapy</td>
<td>2</td>
</tr>
<tr>
<td>89. Low body mass index</td>
<td>2</td>
</tr>
<tr>
<td>90. Lower limb arterial disease</td>
<td>2</td>
</tr>
<tr>
<td>91. Malignancy</td>
<td>2</td>
</tr>
<tr>
<td>92. Mean pulmonary pressure</td>
<td>2, 4, or 5</td>
</tr>
<tr>
<td>93. Metastatic disease</td>
<td>2</td>
</tr>
<tr>
<td>94. Mild liver disease</td>
<td>2</td>
</tr>
<tr>
<td>95. Mitral or aortic valve surgery</td>
<td>6</td>
</tr>
<tr>
<td>96. Mitral surgery</td>
<td>6</td>
</tr>
<tr>
<td>97. Mitral surgery and Pulmonary Artery Pressure</td>
<td>6</td>
</tr>
<tr>
<td>98. Moderate to severe renal disease</td>
<td>2</td>
</tr>
<tr>
<td>99. Morbid obesity ($\geq 1.5$ times ideal weight)</td>
<td>2</td>
</tr>
</tbody>
</table>

*continued next page*
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>100. Multi valve surgery</td>
<td>6</td>
</tr>
<tr>
<td>101. Multivalve or CABG + valve surgery</td>
<td>6</td>
</tr>
<tr>
<td>102. MVR</td>
<td>6</td>
</tr>
<tr>
<td>103. Myocardial infarction</td>
<td>3</td>
</tr>
<tr>
<td>104. Neurological dysfunction</td>
<td>2</td>
</tr>
<tr>
<td>105. No. of endarterectomies</td>
<td>6</td>
</tr>
<tr>
<td>106. Nonuse of IMA</td>
<td>6</td>
</tr>
<tr>
<td>107. Number of diseased coronary vessels</td>
<td>3</td>
</tr>
<tr>
<td>108. NYHA functional classification (I, II, III, IV)</td>
<td>4</td>
</tr>
<tr>
<td>109. One surgeon</td>
<td>6</td>
</tr>
<tr>
<td>110. Operative aortic valve stenosis</td>
<td>6</td>
</tr>
<tr>
<td>111. Operative mitral valve insufficiency</td>
<td>6</td>
</tr>
<tr>
<td>112. Other hydro-electrolytic disturbance</td>
<td>7</td>
</tr>
<tr>
<td>113. Other operation (not CABG or valve)</td>
<td>6</td>
</tr>
<tr>
<td>114. Other rare circumstances (e.g. paraplegia, pacemaker dependency, congenital heart disease in adult, severe asthma)</td>
<td>5</td>
</tr>
<tr>
<td>115. Other significant and uncontrolled systematic disturbance</td>
<td>7</td>
</tr>
<tr>
<td>116. Other than isolated coronary surgery</td>
<td>6</td>
</tr>
<tr>
<td>117. Peripheral vascular disease</td>
<td>2</td>
</tr>
<tr>
<td>118. Poor left ventricular function (ejection fraction &lt; 0.3, or end-diastolic left ventricular pressure above 18 torr at rest)</td>
<td>4</td>
</tr>
<tr>
<td>119. Post myocardial infarction ventricular septal defect</td>
<td>5</td>
</tr>
<tr>
<td>120. Postinfarct septal rupture</td>
<td>5</td>
</tr>
<tr>
<td>121. Preoperative cerebrovascular dysfunction</td>
<td>2</td>
</tr>
<tr>
<td>122. Preoperative intra-aortic balloon pump (IABP)</td>
<td>7</td>
</tr>
<tr>
<td>123. Preoperative intubation</td>
<td>7</td>
</tr>
<tr>
<td>124. Pre-operative therapy with antiplatelet agents</td>
<td>7</td>
</tr>
<tr>
<td>125. Preoperative use of intravenous nitroglycerine</td>
<td>6</td>
</tr>
<tr>
<td>126. Preoperative ventilation</td>
<td>7</td>
</tr>
<tr>
<td>127. Presence of hypertension, diabetes, or peripheral vascular disease</td>
<td>2</td>
</tr>
</tbody>
</table>

*continued next page*
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>128. Pressure Gradient $&gt;120\text{mm Hg}$</td>
<td>4</td>
</tr>
<tr>
<td>129. Previous cardiac surgery</td>
<td>6</td>
</tr>
<tr>
<td>130. Previous coronary bypass surgery</td>
<td>6</td>
</tr>
<tr>
<td>131. Previous myocardial infarction</td>
<td>3</td>
</tr>
<tr>
<td>132. Previous ventricular arrhythmias</td>
<td>5</td>
</tr>
<tr>
<td>133. Prior CABG</td>
<td>6</td>
</tr>
<tr>
<td>134. Prior heart operation</td>
<td>6</td>
</tr>
<tr>
<td>135. Prior vascular surgery</td>
<td>6</td>
</tr>
<tr>
<td>136. PTCA $&lt;6\text{h}$</td>
<td>6</td>
</tr>
<tr>
<td>136. Pulmonary artery systolic pressure $&gt;60\text{mm Hg}$</td>
<td>2, 4, or 5</td>
</tr>
<tr>
<td>138. Pulmonary embolectomy</td>
<td>6</td>
</tr>
<tr>
<td>139. Pulmonary hypertension</td>
<td>2, 4, or 5</td>
</tr>
<tr>
<td>140. Pulmonary rales</td>
<td>4</td>
</tr>
<tr>
<td>141. Recent myocardial infarction $\leq 30\text{d}$*</td>
<td>3</td>
</tr>
<tr>
<td>142. Renal dysfunction</td>
<td>2</td>
</tr>
<tr>
<td>143. Re-operation - any</td>
<td>6</td>
</tr>
<tr>
<td>144. Re-operation - first</td>
<td>6</td>
</tr>
<tr>
<td>145. Re-operation - second or more</td>
<td>6</td>
</tr>
<tr>
<td>146. Resting ST-segment depression</td>
<td>3</td>
</tr>
<tr>
<td>147. Resuscitation</td>
<td>7</td>
</tr>
<tr>
<td>148. Reversible ischemia, unstable angina</td>
<td>3</td>
</tr>
<tr>
<td>149. Saphenous vein graft only</td>
<td>6</td>
</tr>
<tr>
<td>150. Serum Creatinine</td>
<td>2</td>
</tr>
<tr>
<td>151. Serum Creatinine $\geq 141\text{ and }\leq 167\text{ µmol/L (}\geq 1.6\text{ and }\leq 1.8\text{mg/dL)}$</td>
<td>2</td>
</tr>
<tr>
<td>152. Serum Creatinine $\geq 168\text{ µmol/L (}\geq 1.9\text{mg/dL)}$</td>
<td>2</td>
</tr>
<tr>
<td>153. Severe chronic intoxication</td>
<td>2 or 8</td>
</tr>
<tr>
<td>154. Severe hyperlipidaemia/hyperlipidemia</td>
<td>2</td>
</tr>
<tr>
<td>155. Severe neurological disease</td>
<td>2</td>
</tr>
<tr>
<td>156. Severe obesity (Body mass index $&gt;30$)</td>
<td>2</td>
</tr>
</tbody>
</table>

*continued next page*
<table>
<thead>
<tr>
<th>Risk Factor</th>
<th>MF</th>
</tr>
</thead>
<tbody>
<tr>
<td>157. Sex</td>
<td>8</td>
</tr>
<tr>
<td>158. ‘Silent’ ischemia</td>
<td>3</td>
</tr>
<tr>
<td>159. Smoking Status</td>
<td>2</td>
</tr>
<tr>
<td>160. ST changes on preoperative electrocardiogram (ECG)</td>
<td>3 or 7</td>
</tr>
<tr>
<td>161. Surgical priority (elective, urgent or emergent)</td>
<td>6</td>
</tr>
<tr>
<td>162. Symptomatic right heart failure</td>
<td>4</td>
</tr>
<tr>
<td>163. Systemic hypertension</td>
<td>2</td>
</tr>
<tr>
<td>164. Thoracic aortic surgery</td>
<td>6</td>
</tr>
<tr>
<td>165. Transferred admission</td>
<td>8</td>
</tr>
<tr>
<td>166. Transplantation</td>
<td>6</td>
</tr>
<tr>
<td>167. Tricuspid surgery</td>
<td>6</td>
</tr>
<tr>
<td>168. Triglyceride levels</td>
<td>2</td>
</tr>
<tr>
<td>169. Triple vessel disease</td>
<td>3</td>
</tr>
<tr>
<td>170. Type of surgery (CABG only, single valve, complex)</td>
<td>6</td>
</tr>
<tr>
<td>171. Unknown ejection fraction</td>
<td>4</td>
</tr>
<tr>
<td>172. Unstable angina</td>
<td>3</td>
</tr>
<tr>
<td>173. Unstable angina or recent myocardial infarction</td>
<td>3</td>
</tr>
<tr>
<td>174. Use of sequential grafts</td>
<td>6</td>
</tr>
<tr>
<td>175. Valve surgery</td>
<td>6</td>
</tr>
<tr>
<td>176. Ventricular aneurysm</td>
<td>5</td>
</tr>
<tr>
<td>177. Ventricular arrhythmia</td>
<td>5</td>
</tr>
<tr>
<td>178. Ventricular septal rupture</td>
<td>5</td>
</tr>
<tr>
<td>179. Ventricular tachycardia or fibrillation</td>
<td>5</td>
</tr>
<tr>
<td>180. Weight ≤ 65kg</td>
<td>2</td>
</tr>
</tbody>
</table>
Appendix B

Risk Scores for Mortality Following Cardiac Surgery

The following list provides some information about the cardiac surgical scores for the risk of mortality identified by Djamaludin (2006). Each item in the list has five pieces of information. They are (a) the name and abbreviated name of the risk score, (b) citation(s), (c) the definition of the mortality response variable in the training/test data sets, (d) an indication of the procedures undertaken by patients in the training/test data sets, and (e) the number of patients in the training/test data sets.

North American Studies

1. **Score**: A simple classification of risk in cardiac surgery (SIMPLE);
   
   **Reference**: Paiement et al. (1983), Tremblay et al. (1993);
   
   **Response Definition**: Hospital mortality;
   
   **Population**: Cardiac surgical patients from Montréal Heart Institute;
   
   **Study Data**: 500 cases in 1980 and 2,029 cases in 1990.

2. **Score**: Bayesian risk assessment in coronary artery surgery (BAYES);
   
   **Reference**: Edwards and Graeber (1987), Edwards et al. (1988);
   
   **Response Definition**: Death within 30 days of operation;
Population: Patients undergoing CABG surgery with cardiopulmonary bypass at one unnamed hospital;
Study Data: 724 cases from January 1984 to April 1987.

3. Score: The Parsonnet Score (PARS);
   Reference: Parsonnet et al. (1989);
   Response Definition: Any death occurring within 30 days surgery;
   Population: Patients in an “existing database” and from the Newark Beth Israel Medical Center undergoing open-heart surgery procedures;
   Study Data: 3,500 cases from 1982 to 1987 and 1,352 cases from the “past two years” (Parsonnet et al., 1989), respectively.

4. Score: System 97 and Bedside Approximation (SYS97);
   Reference: Bernstein and Parsonnet (2000);
   Response Definition: Surgical mortality defined as death at any time during the same hospital admission;
   Population: Patients undergoing open heart procedures at one of ten New Jersey centres;
   Study Data: Training data set of 8,593 cases from 1994 and 1995.

5. Score: Cardiac Surgery Reporting System, New York State (NYSTATE:90);
   Reference: Hannan et al. (1990);
   Response Definition: In-hospital mortality;
   Population: Adult patients undergoing open heart surgery at one of 28 hospitals in New York State;
   Study Data: 7,596 cases from January to June 1989.

6. Score: Cardiac Surgery Reporting System, New York State (NYSTATE:94);
   Reference: Hannan et al. (1994);
   Response Definition: In-hospital mortality;
   Population: Patients undergoing CABG surgery at one of 30 hospitals in New York State;
7. **Score**: Clinical Severity Scoring System (CSSS);
   
   **Reference**: Higgins et al. (1992);
   
   **Response Definition**: (a) In-hospital mortality regardless of length of stay or mortality within 30 days of hospital discharge, or (b) Severe morbidity as defined by Higgins et al. (1992, Page 2344);
   
   **Population**: Patients in the Cleveland Clinic Foundation database who underwent CABG surgery;
   
   **Study Data**: 5,051 cases from July 1986 to June 1988.

8. **Score**: O’Connor Score (O’CONNOR);
   
   **Reference**: O’Connor et al. (1992);
   
   **Response Definition**: In-hospital mortality;
   
   **Population**: Patients undergoing CABG surgery at one of five clinical centres in northern New England;
   
   **Study Data**: 3,055 cases from July 1987 to 15 April 1989.

9. **Score**: Simplified Clinical Risk Scoring System (SCRSS);
   
   **Reference**: Tuman et al. (1992);
   
   **Response Definition**: (a) Morbidity (including death) as defined by Tuman et al. (1992, Page 37), and (b) Length of stay in intensive care;
   
   **Population**: Patients undergoing cardiac surgery;
   
   **Study Data**: 3,550 cases (time of data collection not stated).

10. **Score**: Bayesian-Logit Model for Risk Assessment (B-LOGIT);
    
    **Reference**: Marshall et al. (1994b);
    
    **Response Definition**: Operative mortality;
    
    **Population**: Patients in the Continuous Improvement in Cardiac Surgery Study, Department of Veteran’s Affairs, undergoing CABG surgery;
    
    **Study Data**: 12,712 cases from April 1987 to March 1990.

11. **Score**: Ontario – Provincial Adult Cardiac Care Network (ONTARIO)
    
    **Reference**: Tu et al. (1995)
    
    **Response Definition**: (a) In-hospital mortality, and (b) post-operative lengths of stay in intensive care and in hospital;
Population: Adult patients undergoing cardiac surgery at one of eight teaching hospitals or one nonteaching hospital in Ontario;

Study Data: 13,098 cases from April 1991 to March 1993.

12. Score: New Ontario – Provincial Adult Cardiac Care Network (ONTARIO:N)
   Reference: Ivanov et al. (1999);
   Response Definition: Operative mortality;
   Population: Patients undergoing CABG surgery undertaken by one of fourteen surgeons from two Toronto teaching hospitals;
   Study Data: 7,491 cases from April 1993 to December 1996.

13. Score: Clinical Risk Score (CRS:MAGOV);
    Reference: Magovern et al. (1996);
    Response Definition: (a) In-hospital mortality, and (b) morbidity arising unexpectedly from any of the major or minor post-operative complications defined by Magovern et al. (1996, Appendix 2);
    Population: Patients undergoing CABG surgery at Allegheny General Hospital, Pittsburgh;
    Study Data: 1,567 cases from July 1991 to December 1992.

14. Score: Society of Thoracic Surgeons Risk Assessment (STS);
    Reference: Edwards et al. (1997);
    Response Definition: Operative mortality;
    Population: Patients in the National Cardiac Surgery Database, Society of Thoracic Surgeons, undergoing CABG surgery;

European Studies

15. Score: The Modified Parsonnet Score (PARS:MOD);
    Reference: Gabrielle et al. (1997);
    Response Definition: Any death within 30 days of surgery;
    Population: Patients undergoing adult cardiac surgery and operated on by
42 of 66 cardiac surgical teams in France;

*Study Data:* 6,649 cases from December 1992 to April 1993.

16. **Score:** French Score (FRENCH);

   *Reference:* Roques et al. (1995);

   *Response Definition:* Any death within 30 days of surgery;

   *Population:* As for PARS:MOD;

   *Study Data:* As for PARS:MOD.

17. **Score:** EuroSCORE (EURO);

   *Reference:* Roques et al. (1999), Nashef et al. (1999);

   *Response Definition:* In-hospital mortality regardless of length of stay or, if the patient is discharged from hospital, death within 30 days of the operation

   *Population:* Patients undergoing cardiac surgery with cardiopulmonary bypass at 128 cardiac surgery centres in eight European countries;

   *Study Data:* 19,030 cases from September to December 1995.

18. **Score:** Patient Specific Predictions (PSP);

   *Reference:* Sergeant et al. (1997);

   *Response Definition:* Survival time following surgery;

   *Population:* Patients undergoing CABG surgery at Katholieke Universiteit (KU) Leuven, Belgium;

   *Study Data:* (a) 5,880 training cases from 1971 to 1987, and (b) 3,720 test cases from July 1987 to January 1992.

19. **Score:** Predictive Risk Model of Catalonia (CATAL);

   *Reference:* Pons et al. (1997);

   *Response Definition:* Surgical mortality defined as death occurring 30 days after the intervention or during hospitalization regardless of length of stay, where hospital-to-hospital transfer is not considered discharge;

   *Population:* Patients over 14 years undergoing open heart surgery at one of the four public and teaching institutions or three private centres in Catalonia. A criterion for including a hospital in the study was that it had undertaken more than 150 procedures during the previous year;
Study Data: 1,309 cases from 14 February 1994 to 31 August 1994.

20. Score: Clinical Risk Score (CRS:STAAT);
   Reference: Staat et al. (1999);
   Response Definition: Morbidity defined as mortality or one of ten non-fatal adverse events given by Staat et al. (1999, Page 961);
   Population: Patients undergoing CABG surgery at one unnamed institution;
   Study Data: 679 cases from January to December 1996.

21. Score: North West England Score (NWENG);
   Reference: Wynne-Jones et al. (2000);
   Response Definition: Operative mortality defined as death within the same hospital admission as operation, regardless of cause. This criterion is met if a patient who was transferred to another hospital died before being discharged.
   Population: Patients undergoing adult cardiac surgery, as defined by Wynne-Jones et al. (2000, Methods), at one of four hospitals in the north-west of England;
   Study Data: 8,210 cases from April 1997 to March 1999.

22. Score: Amphiascore (AMPHIA);
   Reference: Huijskes et al. (2003);
   Response Definition: (a) In-hospital mortality, death during hospitalization at the Amphia hospital or one of the affiliated hospitals; (b) major adverse cardiac events, defined by Huijskes et al. (2003, Section 2.2.2); and (c) extended length of stay, at least three days in intensive care;
   Population: Patients undergoing CABG and/or heart valve operations at the Amphia hospital, Breda, The Netherlands;
   Study Data: 7,282 cases from January 1997 to December 2001.

An Australian Study

23. Score: Australian Society of Cardiac and Thoracic Surgeons Risk Model (ASCTS);
   Reference: Reid et al. (2001);
   Response Definition: Surgical mortality within the period of admission for
surgery;

Population: Patients undergoing cardiac surgical procedures at one of five Victorian cardiac surgical units;

Study Data: 10,753 cases from 1995 to 1999
Appendix C

Studies Evaluating the Parsonnet Score, the EuroSCORE, or both

The following list provides some information about some studies to assess the performance of the Parsonnet Score, the EuroSCORE, or both scores in populations that were remote from the populations which provided the data from which each score was constructed. Each item in the list contains (a) the citation and an abbreviated citation for the study, (b) the score(s) evaluated, (c) the mortality definition used in the study, (d) the patient population which provided the data for the study, and (e) the number of patients in the study dataset and the time they underwent the cardiac surgical procedures.

1. *Study Reference*: Orr et al. (1995) (ORR:95);
   *Score(s)*: Parsonnet score;
   *Response Definition*: Mortality in-hospital or within 30 days;
   *Population*: Patients undergoing CABG surgery at St Vincent Hospital, Worcester, Massachusetts;
   *Study Data*: 868 cases from 1991 to 1993.

2. *Study Reference*: Gabrielle et al. (1997) (GAB:97);
Score(s): Parsonnet score;

Response Definition: Any death within 30 days of surgery;

Population: Patients undergoing adult cardiac surgery and operated on by 42 of 66 cardiac surgical teams in France;

Study Data: 6,649 cases from December 1992 to April 1993.

3. Study Reference: Pliam et al. (1997) (PLI:97);

Score(s): Parsonnet score;

Response Definition: Hospital mortality;

Population: Patients undergoing isolated CABG surgery or combined CABG and valve replacement or repair in the San Francisco Heart Institute (SFHI) Cardiac Interventional and Surgical database;

Study 1 Data: 3,443 cases from January 1991 to June 1994;

Study 2 Data: Subset of 3,237 cases undergoing CABGs only.

4. Study Reference: Weightman et al. (1997) (WEI:97);

Score(s): Parsonnet score;

Response Definition: In-hospital mortality;

Population: Patients undergoing CABG or combined CABG and valve surgery at the Sir Charles Gairdner Hospital, Perth;

Study Data: 927 cases from March 1993 and March 1996.


Score(s): Parsonnet score;

Response Definition: Death within 30 days of operation or in the same hospital admission as operation;

Population: Patients undergoing CABG alone at two hospitals in England;

Study Data: 1,774 cases recruited from the Royal Brompton Hospital, London, April 1994 to March 1995 and from and the Wythenshawe Hospital, Manchester, February 1995 to January 1996.

6. Study Reference: Martínez-Alario et al. (1999) (MAR:99);

Score(s): Parsonnet score;

Response Definition: In-hospital mortality;
Population: Patients undergoing CABG, single and double valve, and congenital heart disease surgery at a tertiary referral centre;

Study Data: 465 cases from January 1997 to May 1998.

7. Study Reference: Geissler et al. (2000) (GEI:00);
Scores: Parsonnet score, EuroSCORE;
Response Definition: Death within 30 days of surgery;
Population: Adult patients undergoing heart surgery with cardiopulmonary bypass at the University of Cologne;
Study Data: 504 cases from September 1998 to February 1999.

8. Study Reference: Peterson et al. (2000) (PET:00);
Score(s): Parsonnet score;
Response Definition: In-hospital mortality;
Population: Medicare patients > 65 years in the Cooperative Cardiovascular Project (CCP) Pilot Revascularization Study undergoing CABG surgery at 28 hospitals in Alabama and Iowa;
Study Data: 3,654 cases from June 1992 to February 1993.

9. Study Reference: Pitkänen et al. (2000) (PIT:00);
Score(s): EuroSCORE;
Response Definition: Death within 30 days of the operation;
Population: Patients undergoing cardiac surgery with cardiopulmonary bypass at Kuopio University Hospital, Finland;
Study 1 Data: 4,592 cases from January 1992 to December 1996;
Study 2 Data: 821 cases from September 1998 and May 1999.

10. Study Reference: Wynne-Jones et al. (2000) (WYN:00);
Score(s): Parsonnet score;
Response Definition: Death within the same hospital admission, regardless of cause;
Population: Patients undergoing adult cardiac surgery, as defined by Wynne-Jones et al. (2000, Methods), at one of four hospitals in the north-west of England;
Study Data: 8,210 cases from April 1997 to March 1999.

11. Study Reference: Kawachi et al. (2001) (KAW:01);
Score(s): Parsonnet score, EuroSCORE;
Response Definition: Death from any cause within 30 days of operation or within the same hospital admission;
Population: Patients undergoing cardiac and thoracic surgery using cardiopulmonary bypass at one of an unknown number of hospitals;
Study Data: 803 cases from August 1994 to December 2000.

12. Study Reference: Nashef et al. (2001) (NAS:01);
Score(s): Parsonnet score;
Response Definition: Death within 30 days or within the same hospital admission as operation;
Population: Patients undergoing open heart surgery at one institution;
Study Data: 6,213 cases from 1996 to 2000.

13. Study Reference: Sergeant et al. (2001) (SER:01);
Score(s): EuroSCORE;
Response Definition: In-hospital mortality, including a stay in a secondary hospital without discharge home;
Population: Patients undergoing primary or repeat CABG surgery of the K.U. Leuven Hospital;
Study Data: 2,051 cases from January 1997 to July 2000.

14. Study Reference: Nashef et al. (2002) (NAS:02);
Score(s): EuroSCORE;
Response Definition: Operative mortality;
Population: Patients in the STS database undergoing coronary or valve surgery;
Study 1 Data: 188,913 cases in 1995;
Study 2 Data: 401,684 cases in 1998/99.

15. Study Reference: Stoica et al. (2002) (STO:02);
Score(s): EuroSCORE;
Response Definition: Hospital death;

Population: Patients undergoing cardiac operations at one institution (in France);

Study 1 Data: 1,575 cases in 1999;
Study 2 Data: 1,543 cases in 2000.

16. Study Reference: Azimakopoulos et al. (2003) (ASA:03);
Score(s): Parsonnet score, EuroSCORE;
Response Definition: Death occurring before hospital discharge;
Population: Patients undergoing CABG surgery with cardiopulmonary bypass at Hammersmith or Harefield Hospitals, London;
Study 1 Data: Parsonnet score, 4,439 cases from 1993 to 1999;
Study 2 Data: EuroSCORE, 4,654 cases from 1993 to 1999.

17. Study Reference: Bridgewater et al. (2003) (BRI:03);
Score(s): EuroSCORE;
Response Definition: Any in-hospital death;
Population: Patients undergoing first-time isolated CABG surgery at one of four hospitals in north-west England;
Study Data: 8,572 cases from April 1999 to March 2002.

18. Study Reference: Huijskes et al. (2003) (HUI:03);
Score(s): EuroSCORE;
Response Definition: Death during hospitalization in the Amphia hospital or one of the affiliated hospitals;
Population: Patients undergoing a CABG and/or heart valve operation at Amphia hospital, Breda, The Netherlands;
Study Data: 7,282 cases from January 1997 to December 2001.

19. Study Reference: Karabulut et al. (2003) (KAR:03);
Score(s): EuroSCORE;
Response Definition: Authors strictly conformed to the definition used by Nashef et al. (1999);
Population: Adult patients undergoing heart surgery with cardiopulmonary
bypass in an unnamed institution;

*Study Data:* 1,123 cases from March 1999 to August 2001.

20. *Study Reference:* Vanagas et al. (2003) (VAN:03);

*Score(s):* Parsonnet score, EuroSCORE;

*Response Definition:* In-hospital mortality;

*Population:* Adult patients undergoing heart surgery with cardiopulmonary bypass at an unnamed institution;

*Study Data:* 444 cases from January to October 2002.


*Scores:* Parsonnet score; EuroSCORE;

*Response Definition:* Death within 30 days of operation;

*Population:* Patients undergoing CABG only surgery, valve procedures with or without CABG, and other miscellaneous procedures at the University Hospital of Lund, Sweden;

*Study Data:* 6,153 cases undergoing 6,222 operations from January 1996 to February 2001.


