14

Risk-Adjusted Monitoring of Outcomes in Health Care

Stefan H. Steiner

University of Waterloo, Waterloo, ON

14.1 Introduction

There is increasing interest among surgeons and health care administrators in monitoring post treatment outcomes, such as hospital length of stay or 30-day mortality, defined as death within 30 days of surgery. Monitoring such outcomes over time allows better oversight of the health care process. In this way serious problems can be rapidly detected and performance can be compared across care providers. In the long run, more quickly eliminating causes of problems and adopting best practices will result in process improvement and better care for patients.

Effective monitoring of health care processes requires a statistical method for two main reasons. First, decisions should incorporate a measure of uncertainty to avoid overreacting to expected outcome variability while still promptly detecting important problems. Here we can borrow ideas developed over the last 80 years for monitoring industrial process outputs. Second, patients (unlike manufactured parts) are not expected to be homogenous and importantly can have dramatically different risks prior to treatment, due for instance to underlying health. Since we want to compare care providers fairly and, for example, do not want to penalize a surgeon who takes on difficult cases, the monitoring should be based on risk-adjusted outcomes.

The goal of this chapter is to outline the main ideas and challenges in risk-adjusted monitoring of health care outcomes. I begin by motivating the need for improvement in health care processes by briefly discussing three recent examples where a lack of proper oversight led to unacceptably high death rates continuing for an extended period of time. Then, in the next section, I introduce the basic concepts behind process monitoring with an industrial example where risk adjustment is not needed. Next, I highlight some of the challenges and unresolved issues in monitoring health care outcomes using a cardiac surgery example to explain the ideas, concerns and methods. The

chapter concludes with a summary and a brief discussion of application areas and possible future extensions of the presented methods.

14.2 Motivation and Background

Medical errors and inefficiencies are a major concern and there is a great need to improve the care of patients. In the landmark report *To Err Is Human: Building a Safer Health System* (Kohn et al., 2000), it was estimated that up to 98,000 preventable deaths are caused by errors in the health system in the United States each year. More recently, Baker et al. (2004) estimated that each year between 9000 and 24,000 deaths in Canadian hospitals are due to mistakes that could have been prevented. Rothschild et al. (2005) concluded that in the intensive care unit, the rates for preventable adverse events and serious errors were 36.2 and 149.7 per 1000 patient-days, respectively.

In light of these problems, it is important to monitor health care outcomes following the provision of medical care or surgical intervention. Such monitoring is also motivated by an increased emphasis on public accountability. In addition, careful monitoring of health care outcomes was recommended by a number of public inquiries conducted after serious problems were identified. The following examples provide illustration.

During 1994, 12 children died as a result of surgery at the paediatric cardiac center at the Winnipeg Health Sciences Centre (Waldie, 1998; Davies, 2001). A subsequent government inquest (Sinclair, 2000) found that many of the deaths were potentially avoidable and that there were systemic problems at the center that could and should have been acted on sooner.

A similar problem occurred at a children's cardiac center in Bristol (England) between 1991 and 1995 (Treasure et al., 1997). The Bristol Royal Infirmary Inquiry (2001) concluded that 30 to 35 children undergoing heart surgery died who probably would have survived if treated elsewhere. The mortality rate at Bristol for cardiac surgery on infants was estimated to be double the rate for the rest of England. The Royal Inquiry made many recommendations for improvement, several of which dealt specifically with "monitoring standards and performance."

In a different context, between 1971 and 1998 the family doctor Harold Shipman killed over 250 mostly elderly patients under his care in England. The inquiry into these events (Smith, 2005) concluded that, among other things, there are "major flaws in the systems that govern death registration, the prescription of drugs and the monitoring of doctors."

These examples portray situations where undesirably high rates of deaths remained undetected, or at least were not acted on, for an undue length of time. This provides motivation for better oversight and the development of methods for monitoring health care outcomes. In such cases, the rapid detec-

tion of poor surgical performance or excess deaths is critical. Early detection of problems would have resulted in prompt investigation of the cause and possibly a criminal investigation in cases of misconduct. This would likely have prevented unnecessary deaths and led to improvement through responses such as procedural changes and retraining.

Statistical methods for monitoring health care outcomes have developed rapidly in recent decades, following earlier work on monitoring industrial processes. I will next briefly describe an example of monitoring an industrial process to highlight the basic ideas of statistical monitoring and then go on to consider some ways that health care outcomes are monitored.

14.3 Monitoring Industrial Processes

To develop methods for monitoring health care outcomes it is natural to build upon the statistical methods for monitoring industrial processes. Beginning with the pioneering work by Shewhart (Shewhart, 1931), over the last 80 years monitoring methods have been further developed and are currently widely used in industry. For an overview of industrial monitoring methods see the Automotive Industry Action Group Reference Manual (AIAG, 1992) and Montgomery (2005).

To illustrate the basic ideas behind statistical monitoring, consider an industrial process that makes oil pans. The oil pan, while not a glamorous product, is critical to the proper functioning of an automobile with an internal combustion engine. Oil pan leaks can lead to catastrophic engine failure if not detected and repaired promptly. Oil pans are produced in a stamping plant by bending and punching pre-produced sheets of metal called blanks. All finished oil pans are inspected and scrapped if any defects such as splits, wrinkles or excessive material thinning are found. To illustrate, a plot is shown in Figure 14.1 of the proportion of oil pans scrapped per day (out of roughly 1125 oil pans) for a 20 day period. The scrap rate varies from day to day due to changes in important process conditions, such as amount of lubricant used, material and geometric properties of the blanks, sharpness of the forming die used to cut and bend the blanks, location of the blank relative to the die in the press, the amount of force used to bend the blank, etc.

To statistically monitor this process, we use the observed variation in the scrap rate during the 20-day period to determine the heights of the three horizontal lines added to Figure 14.1. These lines represent decision limits, derived using statistical models and assumptions, described in Montgomery (2005). The center line represents the average daily scrap rate (.083 over this period). More important are the lower and upper decision limits at heights .058 and .108 respectively, which delineate the expected maximum range of the daily proportion of oil pans scrapped when the process is operating normally.

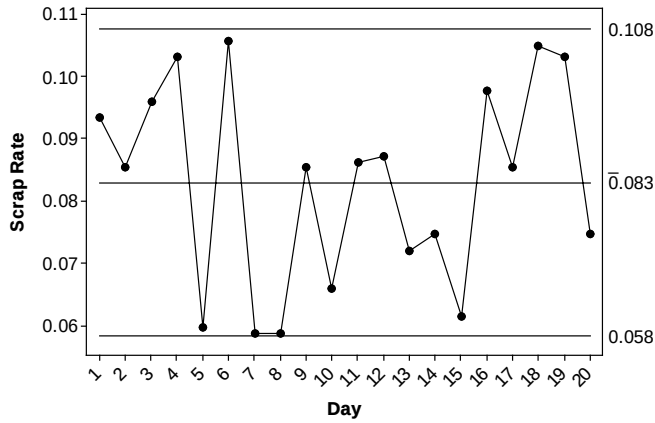

FIGURE 14.1: Chart for monitoring the daily oil pan scrap rate.

If, on the other hand, an unexpected large change in the oil pan production system occurs the scrap rate will fall outside the decision limits and trigger a "signal." For instance, suppose on day 21 the scrap rate is .12 (not shown in Figure 14.1) and thus the chart signals. In this case, based on the decision limits, we believe it is very unlikely that we would see such a large daily scrap rate if the process were operating normally. We therefore would conclude something unexpected happened on day 21 to increase the scrap rate, and would investigate what might have caused this. For instance, suppose we determine that on day 21 we started to use a new batch of blanks and that the specific cause of the signal in this case is that the new batch of blanks had low pliability. Then, to improve the process, one option would be to implement an inspection process to check each new batch of blanks before use, rejecting any with low pliability.

The oil pan example illustrates monitoring pass/fail outputs such as are typical in the health care context. However, many other types of monitoring charts have been developed for different types of outputs and goals (Montgomery, 2005). For instance, when monitoring a continuous output, e.g., the thickness of a part, two charts are often used to monitor the process average and variability separately. Charts for monitoring continuous features are common in industry, since in most cases monitoring a continuous rather than a pass/fail output is preferred due to the more detailed information provided by each observation.

Monitoring charts such as the one displayed in Figure 14.1, which only use data from a single day at a time to make decisions, are good at identifying relatively large process changes, as illustrated by the pliability of the batch of blanks in the example. However, with such charts, smaller, persistent process

changes, such as the slow wearing of a die, will be hard to detect quickly. For smaller persistent changes, cumulative sum (CUSUM) or exponentially weighted moving average (EWMA) charts that accumulate information across multiple time periods are preferred. CUSUM and EWMA charts are commonly used in health care monitoring applications and are described in more detail in the next two sections.

14.4 Monitoring Outcomes in Health Care: Issues and Challenges

Monitoring health care outcomes has some of the same challenges as monitoring industrial processes. We still want rapid detection of problems without overreacting to process variation within acceptable bounds. Quick detection of comparatively poor, or exceptionally good, performance will lead to better public accountability and facilitate improvement by eliminating problems and adopting best practices. However, there are also some important differences between the industrial and health care contexts that we need to consider to successfully monitor health care outcomes.

First of all, there are potential privacy and ethical concerns with the use of health care outcomes data. For instance we need to ensure that any publicly available data cannot be linked to individual people. There is also a human element in responding to signals. If the chart suggests concerns about the performance of a particular surgeon or surgical team we must tread carefully. With any monitoring approach, false alarms can occur due to recent adverse outcomes attributable to bad luck rather than a real change in performance. Also, our monitoring method may be flawed. For example, we could be using an out-dated model for adjusting risk according to the patient mix (see below).

With important health care outcomes we prefer to update the monitoring chart after obtaining results for each individual patient rather than waiting until a fixed time period has passed, e.g. a day, as is typical in industry. In this way we will always be considering the most up-to-date information and have a better chance to quickly detect process changes. However, often health care outcomes are pass/fail, such as 30-day mortality, i.e., whether the patient dies within 30 days of surgery, or the presence or absence of complications. With only a single pass/fail outcome it is not possible to reliably detect process performance changes. As a result, in most health care applications we monitor with CUSUM or EWMA charts that accumulate information over time.

Health care outcomes are also not always available quickly or easy to determine. As an example, the commonly used outcome of 30-day mortality obviously requires 30 days to elapse before it can be observed. In addition, while determining whether a patient has died is clinically straightforward, some patients may be lost to follow-up. For example, imagine a patient survives the

initial surgery and is discharged but then subsequently dies (perhaps after going back to a different hospital). We need to be careful, since without proper follow-up and record keeping we may make mistakes that will negatively affect the performance of the monitoring procedure.

For common health care procedures external standards of performance may be available from clinical research or historical data from other health care providers. Such external standards allow us to compare performance with the standard rather than look for changes in performance in the existing process. In industry external standards are typically not available due to differences in products and competitive pressures. To monitor a health care process by comparing results to an external standard, estimation of the historical process variation is no longer required. Thus, the use of external standards allows rapid development of an appropriate chart. This is a big advantage; however, it alters the nature of the chart and changes our interpretation of what a "signal," i.e., an observation falling outside the decision limits, means. Signals now no longer suggest the process has changed, but rather that we have accumulated enough evidence to conclude that the current process performance is substantially different than the external standard. This has important consequences. It implies that, given a signal, there is no sense in looking for a cause that acted recently since the process may well not have changed but rather that its average performance differs from the external standard. Thus, when comparing to an external standard, process improvement comes from reacting (appropriately) to evidence of significantly better or worse performance than the standard.

Another complication arises because in some health care contexts performance may be continually changing. For example, changes could be due to a learning curve where a novice surgeon improves with practice, or due to rapid innovations in surgical technique. In such cases the only option is to monitor the process by comparing performance to an external standard.

In a health care context it is also common to stratify (group) the outcomes in some way, for example by surgeon or type of surgery, and to compare performance across the strata. Here, the performance of each stratum must be estimated repeatedly as time goes on. Comparisons of this sort are also common in industry, i.e., we might compare machine, operators, production lines, etc. However, few monitoring methods have been developed to meet this goal. Liu et al. (2008) provides an exception.

Finally, and perhaps most importantly, patients do not all have the same risk of an adverse outcome. In industry we assume products are homogeneous and that prior to processing each is equally likely to yield poor results. In health care contexts we must take into account patient characteristics and the possible changing nature of the patient population, since the risk of an adverse outcome prior to treatment depends on numerous patient factors (termed covariates) such as age, gender, underlying health status, etc. Patient mix is the term used to describe the composition of the patient population which contributes to the natural variation of the process. We can not reduce the

variation caused by patient mix in the actual health care process, but we can model the effect of the patient mix prior to medical treatment or surgery and use that model to "risk adjust" the individual observed outcomes. Using risk adjustment to remove the effect of patient mix makes the monitoring method more sensitive to important process changes that we can impact. Risk adjustment is also necessary to allow fair comparisons among surgical or medical care providers or institutions with different patient mixes. Intuitively, risk adjustment is needed because the death of the low risk patient is more indicative of poor performance than the death of a high risk patient, and similarly the survival of a high risk patient is more indicative of good performance than the survival of a low risk patient.

14.5 Monitoring Outcomes in Health Care: Methods

In this section, I briefly discuss health care outcome monitoring that involves some sort of risk adjustment. To illustrate the methods I use a cardiac surgery example presented previously in Steiner et al. (2000) consisting of 6,994 operations from a single UK surgical center over the seven year period 1992–98. The data on each patient includes surgery date, surgeon, type of procedure, surgical outcome and the pre-operative variables which comprise the Parsonnet score (Parsonnet et al., 1989) such as age, gender, hypertension, diabetic status, renal function and left ventricular mass. The Parsonnet score is an established overall measure of risk for adult cardiac surgery. To illustrate the risk-adjusted monitoring methodology we focus on the 30-day post-operative mortality rate. We define

$$y_t = \begin{cases} 1 & \text{if patient } t \text{ dies within 30 days,} \\ 0 & \text{otherwise.} \end{cases}$$

For all risk-adjusted monitoring methods we need to predict the risk of an adverse outcome of each patient prior to surgery. With 461 deaths within 30 days of surgery in these 6994 patients the overall mortality rate was .066. In this example, we could use a mapping of the Parsonnet score to a probability of death within 30 days of the surgery as an external standard. However, due to the large volume of data, I instead build a customized risk model. Following Steiner et al. (2000) and using the data from 1992–93, gives the risk prediction model

$$\log\left(\frac{p_t}{1-p_t}\right) = -3.68 + .077\, x_t, \tag{14.1}$$

where x_t is the Parsonnet score and p_t is the predicted 30-day mortality probability for patient t. This model contains only a single covariate x_t but more

complicated models are possible. For patients between 1992 and 1993 the Parsonnet scores ranged from 0 to 69 and thus the predicted risks of death prior to surgery from (14.1) ranged between .025 and .84.

A number of graphical methods for health care outcome monitoring have been proposed. They all plot some performance summary against time (represented by the patients ordered by surgery date). Here I consider three different approaches. For each approach I use the risk model (14.1) derived from the 1992–93 data and illustrate the proposed prospective monitoring with the patient outcomes for Surgeon 2 starting in 1994. Surgeon 2 left the surgical center in August 1996 and in the period 1994–96 operated on 330 patients. Here we use the internal standard from all surgeons at the center from 1992–93, i.e., the risk model (14.1), as an external standard for Surgeon 2 over the period 1994–96.

14.5.1 Variable Life-Adjusted Display Chart

Lovegrove et al. (1997) and Poloniecki et al. (1998) suggested monitoring using a plot of the difference between the cumulative observed deaths and cumulative predicted deaths given by

$$S_t = \sum_{i=1}^{t} (y_i - p_i), \quad t = 1, 2, 3, \ldots \tag{14.2}$$

That is, S_t is recomputed each time there is a new patient (and their 30-day mortality outcome becomes known). Plots of S_t versus t, known as variable life-adjusted display (VLAD) charts, provide a valuable visual aid where positive values of S_t suggest worse performance than expected. VLAD charts show changes in performance as changes in the slope. However, VLAD charts do not specify how much variation in the plot is expected due to chance under acceptable surgical performance, and hence it is not clear how large a deviation from the horizontal line at zero should raise concern. An example VLAD chart for Surgeon 2 is shown in the top panel of Figure 14.2. We see an increase in the VLAD chart showing that for Surgeon 2 at the end of the series there were more deaths than expected by the risk model (14.1) and the observed patient mix.

14.5.2 Exponentially Weighted Moving Average Chart

An alternative risk-adjusted performance summary for Surgeon 2 is given in the bottom panel of Figure 14.2. The exponentially weighted moving average (EWMA) chart plots M_t versus t, the ordered patient numbers, where we define

$$\begin{aligned} M_t &= \lambda w_t + (1-\lambda) M_{t-1} \\ &= \lambda w_t + (1-\lambda)\lambda w_{t-1} + (1-\lambda)^2 \lambda w_{t-2} + \cdots \end{aligned} \tag{14.3}$$

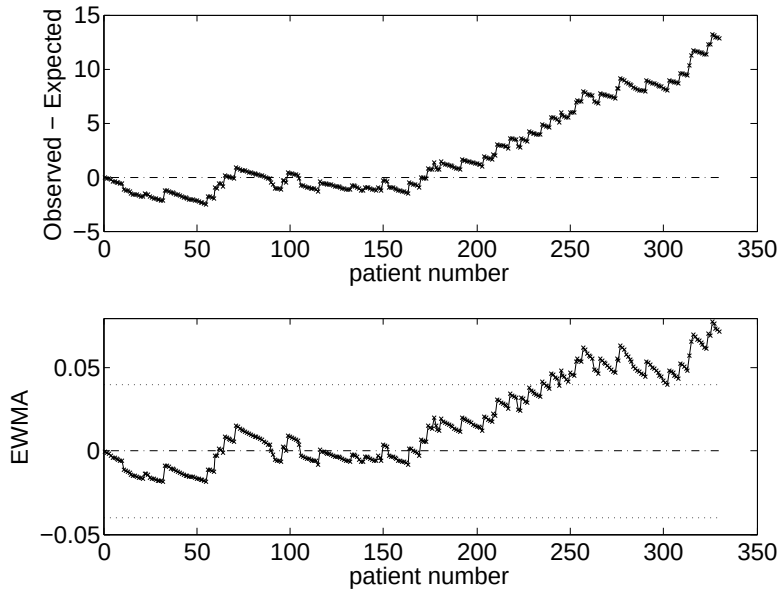

FIGURE 14.2: Risk-adjusted performance summary for Surgeon 2. Top panel: VLAD chart given by (14.2); bottom panel: EWMA chart given by (14.3) with $w_t = y_t - p_t$, $\lambda = .01$ and decision limits of $\pm .04$.

with $0 < \lambda \leq 1$, $M_0 = 0$ and w_t a score given to the tth patient. We can choose appropriate patient scores (w_t) and EWMA smoothing constant (λ) based on the monitoring goals. See Grigg and Spiegelhalter (2007) and Cook et al. (2011) for more details on risk-adjusted EWMA charts.

Here, similar to the VLAD charts in Section 14.5.1, I selected observed minus expected (O–E) patient scores; that is, for the tth patient

$$w_t = y_t - p_t. \tag{14.4}$$

The fifth column of Table 14.1 illustrates how the O–E scores (14.4) reflect surgical performance and provide the risk adjustment. Negative scores reflect good performance with a large negative score ($-.542$) arising from successful surgery on a high risk patient (Parsonnet score = 50). Deaths, on the other hand, contribute a positive score with the death of a high risk patient resulting in a smaller positive score (.458) than the death of a low risk patient (.975).

Since pass/fail outcomes accumulate information slowly, small values of λ are desirable in Equation (14.3), because they allow many patient outcomes to contribute in a substantial way to the EWMA M_t. In Equation (14.3), λ determines how fast the relative weights drop off as patient outcomes are further in the past. For example, with $\lambda = .05$ the score of the most recent

TABLE 14.1: Example patient scores.

Description	Out-come y	Par-sonnet x	Prior Risk p	O–E Scores $y-p$	Likelihood Ratio Scores $R=2$	$R=.5$
Low Risk Success	0	0	.025	−.025	−.024	.012
High Risk Success	0	50	.542	−.542	−.433	.316
Low Risk Death	1	0	.025	.975	.669	−.681
High Risk Death	1	50	.542	.458	.260	−.377

patient has weight .05, the next most recent patient's weight is $.05 \times .95 = .0475$, the patient weight just before that $.05 \times .95 \times .95 = .0451$, etc. In the bottom panel of Figure 14.2 we illustrate the O–E EWMA for Surgeon 2 with $\lambda = .01$, as recommended by Cook (2003). By contrast the VLAD chart in the top panel of Figure 14.2 uses the same patient scores but gives equal weight to all patients. As a result, the EWMA is better than the VLAD chart at detecting recent surgical performance changes. The EWMA also has the advantage that other patient scores, such as (14.6), can be used.

Run length, defined as the time (or number of observations) until a chart signals, can be used to compare the performance of different monitoring charts. When the process is operating as expected we want long run lengths, while if an important change has occurred we want short run lengths to ensure the change is detected quickly. Lucas and Saccucci (1990) provide a method to approximate the EWMA's average run length under different assumptions. Using this approximation I selected decision limits for the EWMA at $\pm.04$ to give an approximate average run length of around 1500 patients when performance matches the predictions from the risk model (14.1). With these decision limits, shown in the bottom panel of Figure 14.2 as horizontal lines, the EWMA first signals evidence that the observed mortality rate for Surgeon 2 is larger than that predicted by the risk model at patient 235. Had this EWMA chart been used prospectively, i.e., as the patient outcomes arose, this signal would have triggered an investigation into the cause and might have resulted in a reaction such as retraining Surgeon 2.

14.5.3 Risk-Adjusted Cumulative Sum Chart

Another alternative monitoring approach, first proposed by Steiner et al. (2000), uses a risk-adjusted cumulative sum (RA-CUSUM) approach based on

$$X_t = \max(0, X_{t-1} + m_t), \tag{14.5}$$

where

$$m_t = \begin{cases} \log\left(\dfrac{1}{1-p_t+Rp_t}\right) & \text{if } y_t = 0, \\ \log\left(\dfrac{R}{1-p_t+Rp_t}\right) & \text{if } y_t = 1, \end{cases} \tag{14.6}$$

$X_0 = 0$ and $R > 1$ is a chosen constant. Similar to the O–E scores (14.4), the patient scores given by (14.6) are positive when a death occurs and negative for a success, as shown in Table 14.1. The cumulative sum given by (14.5) then accumulates evidence of poor performance over time. The cumulative sum X_t of the scores is never allowed to be negative so that a deterioration of performance at any time (even after a series of favorable results immediately before the deterioration) will be quickly detected. The patient scores given by (14.6) use a different scaling and are a good alternative to the O–E scores given in (14.4) since they are optimal (Moustakides, 1986), in terms of average run length, to compare the hypotheses

$$\begin{aligned} \mathcal{H}_0 : \text{odds of death for patient } t &= p_t/(1-p_t), \\ \mathcal{H}_A : \text{odds of death for patient } t &= Rp_t/(1-p_t) \end{aligned}$$

repeatedly over time. Note that under $\mathcal{H}_0$ the odds of death equals what is expected by the risk model (14.1), while $\mathcal{H}_A$ corresponds to a worsening of performance. We choose R based on the size of the process change we are interested in quickly identifying. The change is measured in terms of the odds of death rather than the probability of death for mathematical reasons.

As defined in (14.5) and (14.6), the RA-CUSUM chart tends to increase if performance deteriorates, i.e., the mortality rate increases. To detect mortality rate decreases, i.e., performance improvements, we can also chart

$$Z_t = \min(0, Z_{t-1} - m_t) \tag{14.7}$$

with patient scores (m_t) as in (14.6) but with $R < 1$. Example patient scores based on (14.6) with $R = 2$ (double the odds of death) and $R = .5$ (half of the odds of death) are given in Table 14.1. We see that relative to the O–E scores (14.4) the scores based on (14.6) with $R = 2$ give a smaller positive penalty for low risk or high risk deaths while giving roughly the same (negative) credit for a low risk success. Since the RA-CUSUM scores are based on the optimal likelihood ratio test the RA-CUSUM with $R = 2$ is better, on average, than the O–E EWMA at detecting changes that result in a doubling of the mortality rate.

By using both X_t, given by (14.5), and Z_t, given by (14.7), the RA-CUSUM is sensitive to increases and decreases in the mortality rate just like the VLAD and O–E EWMA charts. Spontaneous process improvements are less likely than performance deterioration but if they occur we want to know as soon as possible, to learn from them and ensure the better performance will continue in the future.

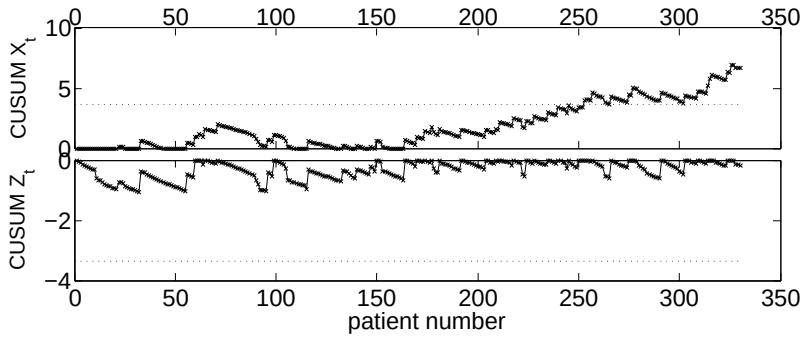

FIGURE 14.3: RA-CUSUM for Surgeon 2. X_t defined by (14.5) with $R = 2$ in top panel with decision limit at 3.65; Z_t defined by (14.7) with $R = .5$ in the bottom panel with decision limit at -3.35.

Since X_t is always non-negative and Z_t is always non-positive we can plot the two RA-CUSUM charts on the same horizontal axis. This is illustrated in Figure 14.3 that shows the RA-CUSUM chart for Surgeon 2 with $R = 2$ in the top half for X_t and with $R = .5$ in the bottom half of Z_t. The RA-CUSUM signals a change in performance if either X_t and Z_t fall outside the chosen decision limits. Setting appropriate decision limits for a RA-CUSUM is discussed in Steiner et al. (2000). In Figure 14.3 the decision limits are shown as horizontal lines at 3.65 and -3.35 for X_t and Z_t respectively and give a combined average run length of roughly 1500 patients (to match the O–E EWMA chart in Figure 14.2) when performance matches the predictions from the risk model (14.1). The RA-CUSUM in Figure 14.3 also shows the increased (risk-adjusted) mortality rate for Surgeon 2 at the end of the series. In this example the RA-CUSUM signals at patient 253; a little later than the O–E EWMA chart given in Figure 14.2.

The performance of the RA-CUSUM chart is explored in Hussein et al. (2011). Grigg and Farewell (2004) provide an overview of risk-adjusted monitoring that includes the RA-CUSUM and some other methods and conclude that in most circumstances the RA-CUSUM is preferred.

14.6 Monitoring Outcomes in Health Care: Uses and Future

I have highlighted here the need for, and uses of, risk-adjusted monitoring of health care outcomes. The three methods presented for process monitoring have proven helpful for controlling and improving health care by providing management oversight of critically important processes. Canadian contributions in this area have been substantial and are ongoing.

Risk-adjusted monitoring has been employed in many clinical settings and different contexts. The popularity of risk-adjusted monitoring is perhaps best exemplified by the publication "Variable Life-Adjusted Displays (VLAD) for Dummies" (Queensland Health's Clinical Practice Improvement Centre, 2008). The VLAD chart has been used extensively. Some examples include monitoring lung surgery outcomes in the Netherlands (Damhuis et al., 2006) and detecting deficiencies in trauma care in England (Tan et al., 2005). The EWMA method has been employed to look for excess deaths in an Australian intensive care unit (Pilcher et al., 2010) among other applications. Specific applications of the RA-CUSUM method include monitoring cardiac surgery outcomes in Canadian hospitals (Harris et al., 2005; Forbes et al., 2005; Novick et al., 2006) and monitoring the length of stay in an intensive care unit in Australia (Cook et al., 2003). In addition, Bottle and Aylin (2008) report the use of RA-CUSUM charts as a management tool to help drive improvement in a number of patient centered outcomes at nearly 100 English hospitals. As well, the RA-CUSUM method was employed to quickly detect problems related to cataract surgery in Western Australia (Ng et al., 2008), excessive radiation doses in Brisbane Australia (Smith et al., 2011) and infectious diseases in foxes in Germany (Höhle et al., 1991).

Research in monitoring health care outcomes is ongoing and appears to be growing. There are a number of important issues and application areas that need further study. In the next few paragraphs I give some examples.

In risk adjustment monitoring not much attention has been given to the goal of comparing performance across surgeons or centers. In the current practice, comparisons are typically only conducted at fixed time points, say every six months (Spiegelhalter, 2005). Methods that would allow comparisons on an ongoing basis as time passes could be valuable since they would allow more timely conclusions and reactions. Here we could use the existing methods to produce separate charts for each surgeon but this would only allow an informal comparison and become unwieldy if there are many surgeons.

An emerging related research area is monitoring public health data where the goal is to quickly detect outbreaks of disease such as influenza. Woodall (2006) and Shmueli and Burkom (2010) provide an overview of the methods and issues. A specific challenge here is the use of nonstandard data sources, such as emergency room visits, non prescription drug purchases, and even the

volume of Internet searches for target key words. Also, the data usually come aggregated in geographical regions and so the best monitoring approaches incorporate the available spatial information. In addition, in most applications there are large numbers of outcomes to monitor. This means we need to be careful to control false alarm rates because too many signals may result in the monitoring charts being ignored. A specific application is given by Google Flu Trends (`www.google.org/flutrends`), which maps global influenza activity by country using aggregated data on Google Internet searches for key words that have been found to correlate with influenza activity.

All risk-adjusted methods for monitoring require specification and fitting of a risk adjustment model. Jones and Steiner (2012) found that the effect of estimation error and model specification error on the performance on the RA-CUSUM chart can be substantial. While their general conclusions are also applicable to other risk-adjusted monitoring approaches, more work is needed. Another related issue is how to best handle monitoring in contexts where we start with little data. Here an important question is when to re-estimate or update the risk model.

This chapter gave an example in which 30-day mortality, a pass/fail health care outcome, was monitored. If it is feasible, working with a continuous outcome is preferred since problems or process changes will then usually be more readily detected. There are many examples in medicine where continuous outcome data are already collected, e.g., hospital lengths of stay, post-surgery survival times, etc. For instance, Biswas and Kalbfleisch (2008), Sego et al. (2009) and Steiner and Jones (2010) all propose risk-adjusted EWMA based methods for monitoring time to death after surgery.

As discussed, monitoring health care processes can lead to improvement if signals are promptly addressed. However, in many applications we should also use more proactive methods that do not wait until we identify trouble before taking action. Lean Six Sigma (De Koning et al., 2006), a general quality improvement methodology that is also borrowed from industry, is at the forefront of such efforts. For more information on improving health care processes I suggest consulting *BMJ Quality and Safety* and the *Journal of Healthcare Quality*. In addition, there are professional organizations devoted to improving healthcare, including the Institute for Healthcare Improvement and the American Society for Quality Healthcare division.

About the Author

Stefan H. Steiner is a professor in the Department of Statistics and Actuarial Science as well as the director of the Business and Industrial Statistics Research Group at the University of Waterloo. He holds a PhD in business

administration (management science/systems) from McMaster University. His research interests include quality improvement, statistical process control, experimental design and measurement system assessment. He is a fellow of the American Society for Quality.

Bibliography

AIAG (1992). *Statistical Process Control Reference Manual*. Chrysler Corporation, Ford Motor Company, and General Motors Corporation.

Baker, G. R., Norton, P. G., Flintoft, V., Blais, R., Brown, A., Cox, J., Etchells, E., Ghali, W. A., Hébert, P., Majumdar, S. R., O'Beirne, M., Palacios-Derflingher, L., Reid, R. J., Sheps, S., and Tamblyn, R. (2004). The Canadian adverse events study: The incidence of adverse events among hospital patients in Canada. *Canadian Medical Association Journal*, 170:1678–1686.

Biswas, P. and Kalbfleisch, J. D. (2008). A risk-adjusted CUSUM in continuous time based on the Cox model. *Statistics in Medicine*, 27:3382–3406.

Bottle, A. and Aylin, P. (2008). Intelligent information: A national system for monitoring clinical performance. *Health Services Research*, 43:1–31.

Cook, D. A. (2003). *The Development of Risk Adjusted Control Charts and Machine Learning Models to Monitor the Mortality Rate of Intensive Care Unit Patients*. Doctoral dissertation, The University of Queensland, Brisbane, Australia.

Cook, D. A., Coory, M., and Webster, R. A. (2011). Exponentially weighted moving average charts to compare observed and expected values for monitoring risk-adjusted hospital indicators. *BMJ Quality and Safety*, 20:469–474.

Cook, D. A., Steiner, S. H., Cook, R. J., and Farewell, V. T. (2003). Monitoring the evolutionary process of quality: Tracking outcomes in intensive care with the risk-adjusted CUSUM. *Critical Care Medicine*, 6:1676–1682.

Damhuis, R., Coonar, A., Plaisier, P., Dankers, M., Bekkers, J., Linklater, K., and Møller, H. (2006). A case-mix model for monitoring of postoperative mortality after surgery for lung cancer. *Lung Cancer*, 51:123–129.

Davies, J. M. (2001). Painful inquiries: Lessons from Winnipeg. *Canadian Medical Association Journal*, 165:1503–1504.

De Koning, H., Verver, J. P. S., van der Heuvel, J., Bisgaard, S., and Does, R. J. M. M. (2006). Lean Six Sigma in healthcare. *Journal for Healthcare Quality*, 28:4–11.

Forbes, T. L., Steiner, S. H., Lawlor, D. K., DeRose, G., and Harris, K. A. (2005). Risk-adjusted analysis of outcomes following elective open abdominal aortic aneurysm repair. *Annals of Vascular Surgery*, 19:142–148.

Grigg, O. and Farewell, V. T. (2004). An overview of risk-adjusted charts. *Journal of the Royal Statistical Society, Series A*, 167:523–539.

Grigg, O. and Spiegelhalter, D. A. (2007). Simple risk-adjusted exponentially weighted moving average. *Journal of the American Statistical Association*, 102:140–152.

Harris, J. R., Forbes, T. L., Steiner, S. H., Lawlor, D. K., DeRose, G., and Harris, K. A. (2005). Risk-adjusted analysis of early mortality following ruptured abdominal aortic aneurysm repair. *Journal of Vascular Surgery*, 42:387–39.

Höhle, M., Paul, M., and Held, L. (1991). Statistical approaches to the monitoring and surveillance of infectious diseases for veterinary public health. *Preventive Veterinary Medicine*, 2009:2–10.

Hussein, A. A., Steiner, S. H., and Gombay, E. (2011). Monitoring binary outcomes using risk-adjusted charts: A comparative study. *Statistics in Medicine*, 30:2815–2826.

Inquiry, B. R. I. (2001). *The Inquiry into the Management of Care of Children Receiving Complex Heart Surgery at the Bristol Royal Infirmary*. Stationery Office, London.

Jones, M. and Steiner, S. H. (2012). Assessing the effect of estimation error on risk-adjusted CUSUM chart performance. *International Journal for Quality in Health Care*, 24:176–181.

Kohn, L. T., Corrigan, J., and Donaldson, M. S. (2000). *To Err Is Human: Building a Safer Health System*. National Academy Press, Washington, DC.

Liu, X., MacKay, R. J., and Steiner, S. H. (2008). Monitoring multiple stream processes. *Quality Engineering*, 20:296–308.

Lovegrove, J., Valencia, O., Treasure, T., Sherlaw-Johnson, C., and Gallivan, S. (1997). Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet*, 18:1128–1130.

Lucas, J. M. and Saccucci, M. S. (1990). Exponentially weighted moving average control schemes: Properties and enhancements. *Technometrics*, 32:1–12.

Montgomery, D. C. (2005). *Introduction to Statistical Quality Control*, Fifth Edition. Wiley, New York.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, 14:1379–1387.

Ng, J. Q., Morlet, N., Franzco, F., Bremner, A. P., Bulsara, M. K., Morton, A. P., and Semmens, J. B. (2008). Techniques to monitor for endophthalmitis and other cataract surgery complications. *Ophthalmology*, 115:3–10.

Novick, R. J., Fox, S. A., Stitt, L. W., Forbes, T. L., and Steiner, S. H. (2006). Direct comparison of risk-adjusted CUSUM and non-risk-adjusted analyses of coronary artery bypass surgery outcomes. *Journal of Thoracic and Cardiovascular Surgery*, 132:386–391.

Parsonnet, V., Dean, D., and Bernstein, A. D. (1989). A method of uniform stratification of risks for evaluating the results of surgery in acquired adult heart disease. *Circulation*, 779:1–12.

Pilcher, D. V., Hoffman, T., Thomas, C., Ernest, D., and Hart, G. K. (2010). Risk-adjusted continuous outcome monitoring with an EWMA chart: Could it have detected excess mortality among intensive care patients at Bundaberg Base Hospital? *Critical Care and Resuscitation*, 12:36–41.

Poloniecki, J., Valencia, O., and Littlejohns, P. (1998). Cumulative risk-adjusted mortality chart for detecting changes in death rate: Observational study of heart surgery. *British Medical Journal*, 316:1697–1700.

Queensland Health's Clinical Practice Improvement Centre (2008). *VLADs for Dummies*. Wiley, Brisbane, Australia.

Rothschild, J. M., Landrigan, C. P., Cronin, J. W., Kaushal, R., Lockley, S. W., Burdick, E., Stone, P., Lilly, C., Katz, J., Czeisler, C. A., and Bates, D. (2005). The critical care safety study: The incidence and nature of adverse events and serious medical errors in intensive care. *Critical Care Medicine*, 33:1694–1700.

Sego, L. H., Reynolds Jr, M. R., and Woodall, W. H. (2009). Risk adjusted monitoring of survival times. *Statistics in Medicine*, 28:1386–1401.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product*. Van Nostrand, New York.

Shmueli, G. and Burkom, H. (2010). Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, 52:39–51.

Sinclair, C. M. (2000). *The Report of the Manitoba Pediatric Cardiac Surgery Inquest: An Inquiry into Twelve Deaths at the Winnipeg Health Sciences Centre in 1994*. Provincial Court of Manitoba, Winnipeg, MB.

Smith, I. R., Foster, K. A., Brighouse, R. D., Cameron, J., and Rivers, J. T. (2011). The role of quantitative feedback in coronary angiography radiation reduction. *International Journal of Quality in Health Care*, 23:342–348.

Smith, J. (2005). *The Shipman Inquiry: The Final Report*. `http://www.shipman-inquiry.org.uk/`.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24:1185–1202.

Steiner, S. H., Cook, R., Farewell, V. T., and Treasure, T. (2000). Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*, 1:441–452.

Steiner, S. H. and Jones, M. (2010). Risk-adjusted survival time monitoring with an updating exponentially weighted moving average (EWMA) control chart. *Statistics in Medicine*, 29:444–454.

Tan, H. B., Cross, S. F., and Goodacre, S. W. (2005). Application of variable adjusted display (VLAD) in early detection of deficiency in trauma care. *Emergency Medicine*, 22:726–728.

Treasure, T., Taylor, K., and Black, N. (1997). Independent review of adult cardiac surgery. Health Care Trust, Unite Bristol, Bristol.

Waldie, P. (1998). Crisis in the cardiac unit. *The Globe and Mail*, October 27 Edition, Section A:3 (column 1).

Woodall, W. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, 38:89–104.