The Monitoring and Improvement of Surgical-Outcome Quality

WILLIAM H. WOODALL

Virginia Tech, Blacksburg, Virginia 24061-0439, USA

SANDY L. FOGEL, MD

Virginia Tech Carilion School of Medicine, Roanoke, Virginia 24014, USA

STEFAN H. STEINER

University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

In this expository paper, we review methods for monitoring medical outcomes with a focus on surgical quality. We discuss the importance and role of risk adjustment. We give the advantages and disadvantages of various competing surveillance methods. We provide an extensive literature review and give some ideas for future research. In addition, we describe the highly effective American College of Surgeons National Surgical Quality Improvement Program (NSQIP), which offers data-based benchmarking of participating hospitals and provides information on best surgical practices. A case study illustrates improvements of mortality and surgical-site infection rates based on the NSQIP approach.

Key Words: CRAM Chart; Cumulative Sum (CUSUM) Chart; NSQIP; Risk-Adjustment; Statistical Process Monitoring; VLAD Chart.

1. Introduction

There has been a great deal of interest in improving the quality of health care, with particular emphasis on surgical quality. The Institute for Healthcare Improvement (IHI) (2014) pointed out, for example, that surgical-site infections continue to represent a significant portion of healthcare-associated infections. Their impact on morbidity, mortality, and cost of care has resulted in their reduction being identified as a top national priority in the U.S. Department of Health and Human Services (2013) "National Action Plan to Prevent Health Care-Associated Infections: Roadmap to Elimination".

According to the healthcare quality framework of Donabedian (1966), one can assess structural, process, or outcome quality. Structural quality refers to the use of metrics such as nurse-to-bed ratios. An example of a process-quality variable would be the percentage of surgical patients receiving antibiotics within a prescribed time period before surgery. Outcome-quality variables, on the other hand, reflect the patients' results. An example of an outcome variable would be whether or not a surgical patient developed a surgical-site infection within 30 days following surgery.

Generally, the proper use of outcome variables requires considerably more effort in data collection, but it is the most informative approach. Porter and Teisberg (2007) and Department of Health (2010), among others, have made a strong case for the use of outcome results to provide vital feedback on what works and what does not instead of concentrating on pro-

Dr. Woodall is Professor in the Department of Statistics at Virginia Tech. He is a Fellow of ASQ. His email address is bwoodall@vt.edu.

Dr. Fogel is NSQIP Surgeon Champion at the Carilion Clinic. His email address is slfogel@carilionclinic.org.

Dr. Steiner is Professor in the Department of Statistics and Actuarial Science at the University of Waterloo. He is a Fellow of ASQ. His email address is shsteine@uwaterloo.ca.

cess targets. The Donabedian (1966) framework was outlined in more detail in the excellent paper by Ko (2009), who also stressed the importance of using outcome results.

Overviews of process monitoring in healthcare applications have been provided by Benneyan (1998a, b), Woodall (2006), Winkel and Zhang (2007, 2012), Woodall et al. (2012), and others. In our paper, we restrict attention to the various approaches for monitoring surgical-outcome quality. Blackstone (2004) reviewed the history of process monitoring in surgical applications and gave an overview of some of the important issues.

Because effective monitoring requires one to account for the variation among patients, we briefly review risk adjustment in Section 2. In Section 3, we give our notation and define some of the surveillance-method performance metrics. In Section 4, we describe the various surveillance methods. In Section 5, we discuss the related analysis of surgical learning curves. Outcome monitoring is important and useful in detecting and understanding changes in performance. It can motivate the need for improvement, which can then come through targeting the most promising opportunities and implementing improvement projects. For this reason, we provide an overview of the American College of Surgeons National Surgical Quality Improvement Program (NSQIP) in Section 6 with an NSQIP-based case study given in Section 7. Some ideas for future research on outcome monitoring are given in Section 8 and our conclusions follow in Section 9.

2. Risk Adjustment

Surgical patients vary considerably with respect to physical characteristics, such as age and weight and with respect to health status. In order to perform useful monitoring or meaningful comparisons of the surgical outcomes stratified by surgeons or hospitals, there must be adjustments for the patient mix of risk factors. Many aspects of risk adjustment were covered in considerable detail by Iezzoni (2012).

Most often logistic regression models are used for patient-level risk adjustment when binary outcomes are used. See, for example, Cohen et al. (2009) and New York State Department of Health (2001). The probability of a particular adverse event, such as death within 30 days of surgery, is modeled with various physical and health characteristics used as the explanatory variables. Examples would be the use of scores, such as the Parsonnet score or the EuroSCORE II, in a logistic regression model for risk adjustment for adult cardiac surgery (Parsonnet et al. (1989), Nashef et al. (2012)). Risk factors used to calculate these types of scores can include gender, age, diabetic status, hypertension status, dialysis status, and so forth.

In virtually all cases, the logistic regression models are based on discrete or categorical explanatory variables modeled using indicator variables and do not include interaction terms. The number of explanatory variables is often in the range from 20 to 30. Models have been built for many other types of adverse events such as surgical-wound infection, anastomotic leak (a leak at the surgical connection of two structures), and deep vein thrombosis (Bruce et al. (2001)).

The risk-adjustment model must be fit based on some historical data from all of the surgeons or hospitals of interest. Most often, the data from a particular time period is somewhat arbitrarily selected to serve as the baseline. Paynabar et al. (2012) discussed the analysis of historical baseline data.

Cook et al. (2008) provided an excellent discussion of some of the issues related to risk adjustment. It is important that the risk-adjustment model accurately estimate the probability of the adverse event of interest because the risk-adjusted surveillance methods detect deviations from the risk-adjustment model predictions. It is also important that risk-adjustment models be periodically updated because they can begin to overestimate the probability of the adverse event due to process improvement.

In comparing hospital performance, it is common to use random-intercept multilevel logistic regression modeling or other types of hierarchical generalized linear models. See, for example, Clark et al. (2010) and COPPS-CMS White Paper Committee (2012).

3. Monitoring Background

It is important to note that in industrial applications the monitoring methods are designed based on information obtained using background data collected from the particular process of interest. The collection and analysis of these data are referred to as phase I, an area reviewed by Jones-Farmer et al. (2014). In most of the methods covered in our paper, however, the performance of a particular hospital or surgeon is monitored relative to a risk-adjustment model constructed using data from a number of hos-

pitals. Thus, it is often performance to a standard that is being monitored. Steiner (2014) discussed this issue in more detail.

We let p_{0i} represent the probability obtained from the risk-adjustment model that the ith surgical patient, $i = 1, 2, \ldots$, experiences the adverse event of interest. We let $Y_i = 1$ if the ith patient experiences the adverse event of interest and $Y_i = 0$ otherwise. The assumption that Y_i, $i = 1, 2, 3, \ldots$, are mutually independent is ubiquitous. The odds against the adverse event occurring are p_{0i}: $1 - p_{0i}$ under the risk-adjustment model. A change in the odds ratio of size δ leads to odds against the event of δp_{0i}: $1 - p_{0i}$ with a corresponding probability of the event of $p_{1i} = \delta p_{0i}/(1 + (p_{0i}(\delta - 1))$. Monitoring methods can be designed to detect specified changes in the odds ratio.

Risk-adjusted monitoring methods are usually compared based on the average run length (ARL), where the run length is the number of surgical patients until a signal is given that there has been a change in the process. We would like the ARL when the process is stable, i.e., the in-control ARL, to be large and the ARL to be small when there is a significant change in the odds of the adverse event.

The ARL can be calculated assuming that any process shift occurred before monitoring begins (called the zero-state ARL) or assuming any shift occurs sometime after the start of monitoring (steady-state ARL). We prefer the use of the steady-state ARL because the assumption of a possibly delayed change in the process seems more realistic. Gombay et al. (2011) pointed out, however, that the in-control ARL can be misleading and considered other metrics, such as the probability of a false alarm within a given number of patients. Sun and Kalbfleisch (2013) also took this latter approach.

4. Various Monitoring Methods

Grigg and Farewell (2004a), Rogers et al. (2004), and Cook et al. (2008) provided review papers on risk-adjusted monitoring. Cook et al. (2008) provided an appendix with all related formulas. Woodall (2006) included a section on risk-adjusted monitoring. A considerable amount of research has been done in the last decade or so, however, with many more applications. A nontechnical review and discussion of issues related to risk-adjusted monitoring was given by Steiner (2014).

Monitoring can lead to insights and is useful for detecting changes in performance and understanding trends over time. Outcome monitoring can be used to identify problems, motivate the need for improvement, and quantify the extent to which improvement initiatives have been successful.

4.1. Risk-Adjusted Sets Method and Resetting SPRT

In their review, Grigg and Farewell (2004a) focused to some extent on the risk-adjusted sets method of Grigg and Farewell (2004b), which has not become widely used. To signal a process deterioration, the sets method requires that the waiting time between adverse events, measured in the number of cases, be below a specified threshold for a specified number of consecutive adverse events.

Sego et al. (2008) showed, in the non-risk-adjusted case, that the apparent performance advantage of the sets method is due to an implicit "headstart" feature that leads to good zero-state ARL performance, but poor steady-state ARL performance relative to competing methods. These performance results likely carry over to the risk-adjusted application.

We do not recommend the use of the resetting risk-adjusted sequential probability ratio test (RSPRT) proposed by Spiegelhalter et al. (2003) and Grigg et al. (2003). This method was discussed in some detail by Cook et al. (2008) and applied by Rogers et al. (2005). With this approach, one sets up a sequential probability-ratio hypothesis test (SPRT) with the null hypothesis corresponding to the risk-adjustment model being correct and the alternative hypothesis corresponding to a specified shift in the odds of the adverse event occurring. If the null hypothesis is accepted, then the hypothesis test is repeated. Rejecting the null hypothesis is a signal that performance may have changed.

There are two issues with this approach. The first is that the SPRT type I and type II error probabilities, α and β, respectively, are often misinterpreted and are not meaningful in assessing the run-length performance of the RSPRT. The second issue is that the RSPRT is a generalization of the risk-adjusted Bernoulli cumulative sum (RA-CUSUM) method discussed in Section 4.4 with the primary effect being that the RSPRT chart is generally less able to detect deterioration in performance after a period of good performance and vice versa. This phenomenon is referred to a building up "credit" in the healthcare surveillance literature and referred to as issues

with "inertia" in the industrial statistical process-monitoring literature.

4.2. Risk-Adjusted p Chart

Alemi et al. (1996) and Alemi and Oliver (2001) proposed aggregating the patients into consecutive groups and using the mean and variance of the sum of the Bernoulli observations within each group to determine Shewhart-type control limits. Hart et al. (2003, 2004) gave related work on Shewhart charts based on aggregated data. Cockings et al. (2006) gave examples of risk-adjusted p charts with patients grouped into consecutive blocks of size 30. Gustafson (2000), on the other hand, compared various types of charts for monitoring infection rates with data aggregated by month.

Although these approaches can provide a good overview summary of the performance of the process over time, greater levels of aggregation cause longer delays in detecting changes in the process relative to methods that incorporate the data on a case-by-case basis. This point was also made by Cook et al. (2003). The adverse effect of data aggregation on monitoring was discussed in a general context by Schuh et al. (2013).

4.3. CRAM and VLAD Charts

The popular variable life adjusted display (VLAD) method of Lovegrove et al. (1997) and the equivalent cumulative risk-adjusted mortality (CRAM) method of Poloniecki et al. (1998) are based on plots of cumulative sums of either $p_{0i} - Y_i$ or $Y_i - p_{0i}$. In the first case, the Y-axis is often labeled "Lives Saved" or "Statistical Lives Saved", whereas, in the second case, it is frequently labeled "Excess Mortality". Sometimes the plot is referred to as an observed minus expected (O-E) CUSUM chart. The VLAD name is most common, however, so we will use it in our paper.

The book by the Clinical Practice Improvement Centre (2008) provides a detailed description of the use of the VLAD and many practical issues related to its use. Albert et al. (2003) and Lovegrove et al. (1999) provided a number of examples of VLAD plots. Treasure et al. (2004) and Sherlaw-Johnson et al. (2000) also discussed the use of the VLAD chart.

The VLAD chart shown in Figure 1 was provided to us by Dr. Albert Yuen of the Hong Kong Hospital Authority. It shows the net lives saved/lost following emergency operations at a hospital in Hong

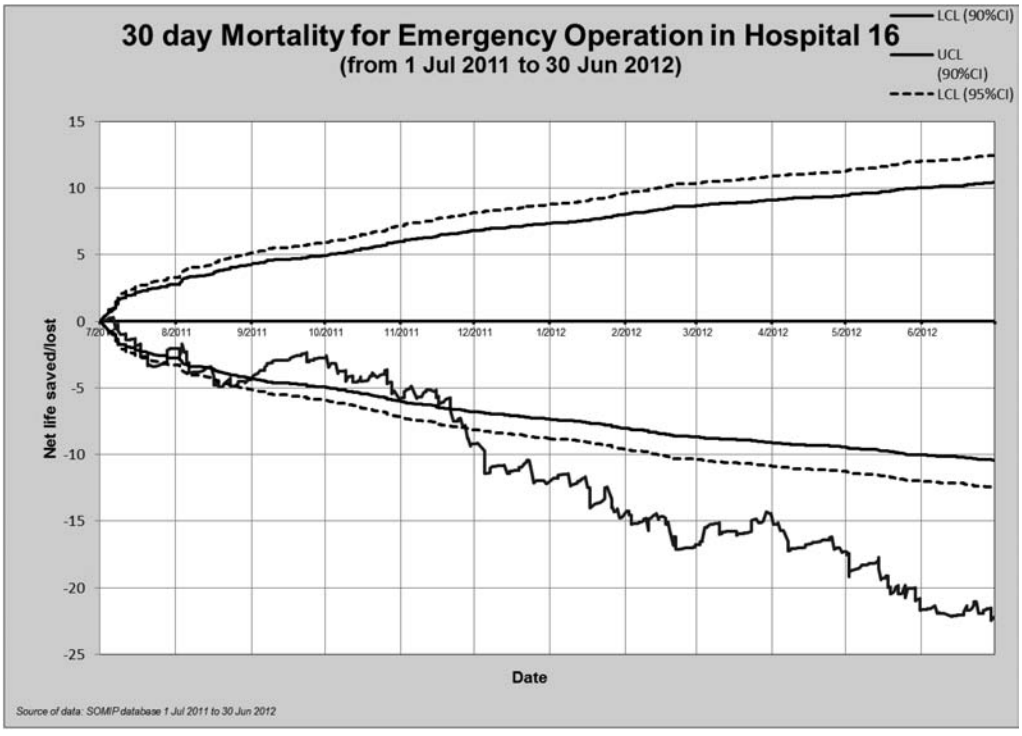

FIGURE 1. Rocket Tail VLAD Plot. Provided by W.-C. Yuen, Hong Kong Hospital Authority.

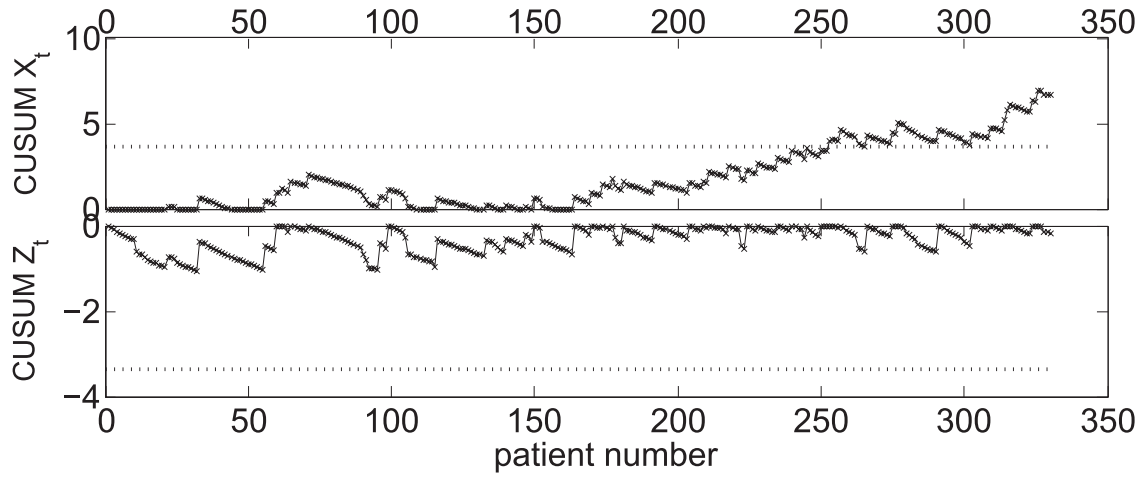

FIGURE 2. Risk-Adjusted Bernoulli CUSUM Chart for Surgeon 2 (from Steiner, (2014)).

Kong. Under the risk-adjustment model, the VLAD statistic wanders over time in a nonstationary manner. It has no tendency to return to any particular value, including zero, because it can be modeled as a random-walk process. The statistics in Figure 1 decrease substantially, however, over time, indicating deterioration in performance compared with the risk-adjustment model. The "rocket-tail" control limits in Figure 1, which widen over time, are based on percentiles of the marginal distribution of the cumulative sum. These limits have been recommended, for example, by Grunkemeier et al. (2003, 2009) and Noyez (2009). The use of these limits to signal changes in quality leads to a problem with inertia because process deterioration, for example, could occur when the VLAD statistic is near the upper limit. In addition, the percentile values used to determine the limits are not directly related to the run-length performance of the method.

Sismanidis et al. (2003) and Poloniecki et al. (2003) proposed an ad hoc signal rule for the VLAD and CRAM charts. The lack of a theoretically justifiable way to determine a signal rule for the chart has led, however, to the use of the VLAD as an easily understood visual aid with reliance on the RA-CUSUM chart, as described in the next subsection, to signal shifts in the performance of the surgical process.

4.4. Risk-Adjusted Bernoulli CUSUM Chart

The risk-adjusted Bernoulli cumulative sum (CUSUM) chart (referred to as the RA-CUSUM chart) of Steiner et al. (2000, 2001) is preferred over the VLAD chart for detecting changes in performance. The RA-CUSUM chart is a generalization of the Bernoulli CUSUM chart of Reynolds and Stoumbos (1999), which was used by Leandro et al. (2005) to monitor the outcomes of liver-transplant surgery.

The one-sided RA-CUSUM chart statistics for a chart designed to detect a change in the odds ratio to a specified value are

$$S_i = \max(0, S_{i-1} + W_i), \quad i = 1, 2, 3, \ldots,$$

where $S_0 = 0$, $W_i = ln(p_{1i}/p_{0i})$ if $Y_i = 1$ and $W_i = \ln[(1-p_{1i})/(1-p_{0i})]$ if $Y_i = 0$. The chart signals when $S_i > h$, where h is selected to provide a specified in-control ARL. The selection of h depends on the relevant population of risk scores.

Often the RA-CUSUM chart is implemented to be two-sided with the simultaneous application of two one-sided charts, one to detect improvement in performance and the other to detect deterioration. The signs on the statistics on the chart used to detect improvement are usually changed so the two one-sided charts can be plotted together more easily. An example of a two-sided RA-CUSUM chart provided in Steiner (2014) is given in Figure 2. In this example, the outcome is mortality within 30 days of cardiac surgery with risk-adjustment based on logistic regression using Parsonnet scores. The upper CUSUM chart was designed to detect a doubling of the odds ratio while the lower CUSUM chart was designed to detect a halving of the odds ratio. The upper CUSUM boundary was crossed with patient 253, indicating poor performance relative to the risk-adjustment model.

There have been a number of applications of the RA-CUSUM chart in the literature. For example,

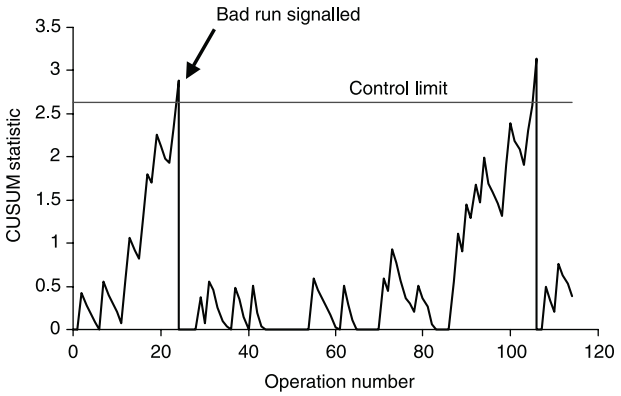

FIGURE 3. RA-CUSUM Chart for Surgeon A with Two Signals of Poor Performance. (Reprinted by permission from Macmillan Publishers Ltd: *Journal of the Operational Research Society*, Sherlaw-Johnson et al., 2007, published by Palgrave Macmillan.)

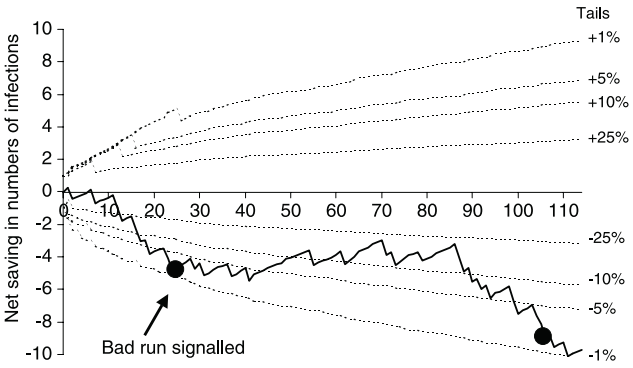

FIGURE 4. Rocket-Tail VLAD Plot for Surgeon A with Two RA-CUSUM Chart Signals Indicated. (Reprinted by permission from Macmillan Publishers Ltd: *Journal of the Operational Research Society*, Sherlaw-Johnson et al., 2007, published by Palgrave Macmillan.)

Axelrod et al. (2006, 2009) discussed the use of the RA-CUSUM method to assess the performance of organ-transplant centers. Beiles and Morton (2004) and Collins et al. (2011) gave applications in assessing the performance of arterial surgery and gastro-esophageal surgery, respectively. Morton et al. (2008) discussed the monitoring of healthcare-acquired infections and used the RA-CUSUM chart as an example. Bottle and Aylin (2008) discussed the reliance on RA-CUSUM charts in a system for monitoring clinical performance involving 100 hospitals in England. For other applications, see Harris et al. (2005), Novick et al. (2006), Moore et al. (2007), and Chen et al. (2011).

It has been recommended that one display the more easily interpretable VLAD chart, but with a RA-CUSUM chart run in the background to signal any changes in quality. This approach was advocated by Sherlaw-Johnson (2005), Sherlaw-Johnson et al. (2005, 2007), Cook et al. (2008), Clinical Practice Improvement Centre (2008), and Collett et al. (2009). We also support this approach, which is illustrated in Figures 3 and 4.

A number of researchers have studied the performance of the RA-CUSUM chart. Jones and Steiner (2012) investigated the effect of estimation error on the performance of the chart. Webster and Pettitt (2007) studied some technical issues related to the computation of chart performance metrics.

Steiner et al. (2001) and Tian et al. (2015) showed that, for a given risk-adjustment model and given control limits, the in-control ARL depends heavily on the population of risk scores. They showed that the effect was considerably greater than that found by Loke and Gan (2012). Zhang and Woodall (2015), however, have developed a method for designing the RA-CUSUM charts based on the method of Shen et al. (2013). This is a computationally intensive approach which requires on-line dynamic simulation to determine the control limits patient-by-patient to design each chart for the specific sequence of patient risk scores observed. This means each chart is customized to the specific sequence of patients at hand. This alleviates the major disadvantage of the risk adjusted Bernoulli CUSUM chart, which is any concern that misspecification or changes in the risk score population can affect to a considerable extent the in-control performance of the chart.

It is frequently stated that the RA-CUSUM has optimal performance based on the results of Moustakides (1986), but these optimality results are based on the assumption of an independent and identically distributed sequence of observations. The observations in the sequence Y_i, $i = 1, 2, 3, \ldots$ are assumed to be independent but they are not identically distributed because patients have varying risk factors.

Grigg and Farewell (2004) pointed out the usefulness of the approximation

$$(\mathrm{ARL_C})^{-1} \cong (\mathrm{ARL_L})^{-1} + (\mathrm{ARL_U})^{-1},$$

where $\mathrm{ARL_C}$ is the ARL for the two-sided RA-CUSUM chart and $\mathrm{ARL_L}$ and $\mathrm{ARL_U}$ are the ARLs for the component lower- and upper-sided RA-

CUSUM charts, respectively. Megahed et al. (2011) gave conditions under which this relationship is exact.

4.5. Risk-Adjusted Exponentially Weighted Moving-Average Methods

Cook at al. (2011) proposed a risk-adjusted exponentially weighted moving average (RA-EWMA) method. The advantages given for their RA-EWMA method are that it communicates information about the current level of an indicator in a direct and understandable way and it explicitly displays information about the current patient case mix. Also, because it is not reset after a signal, they considered the RA-EWMA chart to be a more natural chart to use in healthcare applications, where a process of care can rarely be changed quickly. One might note that it is common not to reset surveillance statistics to their initial values after a signal in prospective public health surveillance applications.

Steiner and MacKay (2014) also proposed an EWMA-based approach that gives more weight to recent outcomes and plots a clinically interpretable estimate of the failure rate for a "standard" patient. They pointed out some advantages of their EWMA approach compared with those of Cook et al. (2011) and Grigg and Spiegelhalter (2007). One advantage claimed is that fewer historical data are needed to set up their surveillance method.

Regardless of any relative advantages of these three EWMA methods, the RA-CUSUM chart is likely to remain the most accepted surveillance approach for monitoring with risk-adjusted binary outcomes in the near future.

4.6. Methods Based on Survival Models

There has been more recent work that incorporates information on the times of any deaths within the given time window following surgery or survival times more generally. Monitoring of survival times is more common in organ-transplantation applications. For reviews of methods used to assess the performance of organ-transplantation centers and related issues, we recommend Collett et al. (2009) and Neuberger et al. (2010). For mortality outcome data in organ-transplantation applications, the time window after surgery is most often one year.

Survival model-based surveillance methods have been proposed by Biwas and Kalbfleisch (2008), Sego et al. (2009), Steiner and Jones (2010), Gandy et al. (2010), and Sun and Kalbfleisch (2013). These methods can lead to better statistical performance than the RA-CUSUM chart based on binary outcomes, but they are considerably more complicated and require a survival model assumption. Phiniketttos and Gandy (2014), however, recently proposed a nonparametric approach based on the Kaplan–Meier survival curve estimator.

Steiner and Jones (2010) showed that the method of Sego et al. (2008) is at a performance disadvantage relative to the competing methods. This is because, until information on the final status of a particular patient is available, no information on patients operated on after this patient can be used in the analysis. In recent developments, Snyder et al. (2014) discussed the application of the CUSUM method of Sun and Kalbfleisch (2008) in the monitoring of survival times after organ transplantation. In addition, Assareh and Mengersen (2014 a, b) proposed Bayesian methods for change-point estimation for step shifts and trends, respectively, in monitoring risk-adjusted survival times.

4.7. Other Approaches

There have been a number of alternative approaches proposed for the monitoring of risk-adjusted binary outcomes. Steiner et al. (1999) proposed methods for monitoring paired binary surgical outcomes, which allows the simultaneous monitoring of mortality and "near-misses". Chang (2008) compared a risk-adjusted Shiryayev–Roberts scheme to the performance of the RA-CUSUM chart. The Shiryayev–Roberts method was found to be less able to detect deterioration in performance.

Gan and Tan (2010) proposed a risk-adjusted version of the time-between events chart, which in the non-risk-adjusted application typically involves use of the geometric distribution. Albers (2011) proposed a related method based on the number of cases between a specified number of adverse events, which led to a generalization of the negative binomial distribution. In the non-risk-adjusted case, however, Szarka and Woodall (2011) reported that these types of time-between-event charts fare poorly compared with the performance of a CUSUM chart.

Zeng and Zhou (2011) proposed a risk-adjusted monitoring method based on Bayesian methods that is said to require less data than other methods in order for monitoring to begin. More recently, Gan et al. (2012) proposed a generalization of the RA-CUSUM chart to detect changes in parameters other than the odds ratio.

5. Use of a Sequential Technique to Assess Learning Curves

It is important to distinguish between the use of the RA-CUSUM method and the use of what is referred to as the CUSUM technique to assess surgical learning curves. With learning curves, the cumulative number of failures is often plotted against the patient number if there is no risk adjustment. Sometimes cumulative values of $Y_i - p_0$ are plotted, where p_0 is the acceptable or expected failure rate. Decision lines are determined based on the sequential hypothesis testing (SPRT) approach of Wald (1947) for a sequence of independent Bernoulli outcomes, where one must also specify an unacceptable failure rate and type I and type II error rates. Rogers et al. (2004) referred to this method as resulting in a "cumulative failure chart". The method can be risk adjusted. The SPRT stopping rule, under which sampling stops when a decision line is crossed, is typically not followed.

For more information on this CUSUM learning curve technique, see Williams et al. (1992), Novick and Stitt (1999), Bolsin and Colson (2000), Novick et al. (2001), Grunkemeier et al. (2003), and Yap et al. (2007). Most of the CUSUM applications found in the review by Biau et al. (2007) were learning curve analyses. It is frequently incorrectly stated in the literature that the learning curve CUSUM is related to the CUSUM method of Page (1954), whereas it is the RA-CUSUM method that is an extension of Page's work. It seems that the learning curve SPRT approach has been misleadingly referred to as a CUSUM procedure because this was the terminology used by de Leval et al. (1994).

In applications of the learning curve approach, it is frequently expected that performance levels will change over time, perhaps more than once. Performance could either improve or deteriorate. Thus, it does not seem reasonable to use a method relying on a one-sided hypothesis testing approach where the error probabilities are based on the assumption of a constant level of performance. In learning curve applications, it is assumed that there is a sequence of independent Bernoulli observations where the interest is in detecting changes in the probability of failure. The large literature on this topic was reviewed by Szarka and Woodall (2011).

6. National Surgical Quality-Improvement Program

Over 560 U.S. hospitals and health systems and 43 outside the U.S. are American College of Surgeons National Surgical Quality Improvement Program (NSQIP) participants. Maggard-Gibbons (2013) reported that about 10% of the hospitals in the U.S. participate in NSQIP and that they account for about 30% of the over 40 million operations performed annually. The participating hospitals provide to NSQIP data on samples of surgical patients designed to represent all types of surgical procedures and all surgeons. NSQIP then provides risk-adjustment models and performance results to the hospitals. Thus hospital administrators and surgeons can see how their results compare to other hospitals and how individual surgeons compare to each other and to the overall NSQIP performance. This process can be used to identify areas needing improvement. Without the sort of benchmarking provided by NSQIP, hospital administrators and surgeons cannot accurately assess their current performance or as easily identify areas most needing improvement. Maggard-Gibbons (2013) reported that in a survey of NSQIP participants, over half reported that prior to joining NSQIP they did not know their surgical mortality rates, much less how their rates compared to other hospitals.

Ko (2009) and Maggard-Gibbons (2013) provided very good descriptions of the history and structure of NSQIP. Also, see http://site.acsnsqip.org/. Cohen et al. (2013) discussed a number of statistical aspects of the NSQIP analyses. It is reported at the NSQIP website that hospitals participating in NSQIP benefit from an average savings of about $3 million per year, reduced readmissions and lengths of stay, higher patient satisfaction, better patient outcomes, better performance on publicly reported measures, and better performance under pay-for-performance programs. In a thorough study, Hall et al. (2009) found that surgical outcomes improved across a majority of the NSQIP participating hospitals in the private sector. Improvement was found for both poor- and well-performing facilities. They reported that NSQIP hospitals appeared to be avoiding substantial numbers of complications, improving care, and reducing costs.

A key aspect leading to the success of NSQIP is that each hospital has a surgeon champion that identifies and leads improvement initiatives. There is also a well-trained surgical clinical reviewer, who is responsible for collecting complete and accurate clinical data, as opposed to reliance on less accurate and less relevant claims or administrative data. About 140 variables are measured for each surgical

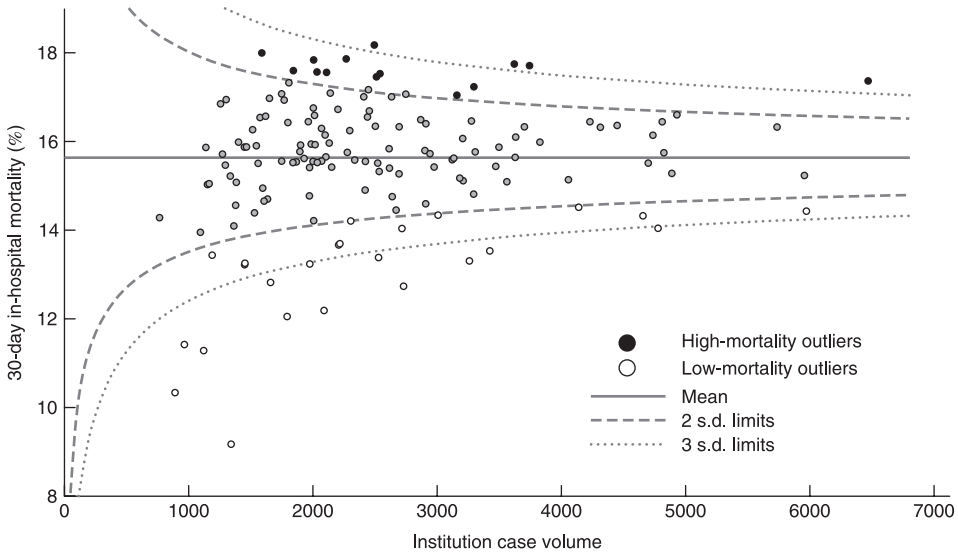

FIGURE 5. Funnel Plot Showing All-Cause Risk-Adjusted In-Hospital 30-Day Mortality for English National Health Service Hospital Trusts (from Symons et al., 2013). Reprinted with permission from John Wiley and Sons.

patient included in the NSQIP database. There are 40 adverse events for which risk-adjustment models are constructed. These include such adverse events as cardiac occurrences, pneumonia, unplanned intubation, ventilator dependence over 48 hours, renal failure, and urinary tract infection, as well as death.

NSQIP previously used "caterpillar" plots to identify outlying performance. The healthcare providers were ordered by the O/E ratio, i.e., the ratio of the observed number of adverse events during a given time period to the expected number based on the risk-adjustment model. Note that, if the adverse event is death, the O/E ratio is often referred to as the standardized mortality ratio (SMR). If the confidence interval on the O/E ratio falls completely below unity, then the hospital is considered to have better than expected performance. A confidence interval falling completely above unity indicates performance below expected. Because much of the ordering of the performance of the hospitals is random, this caterpillar plot can lead to an overemphasis on relative position. Small differences in performance can be attributed to chance and small differences can change the position of a hospital on the caterpillar plot substantially. For this reason, we prefer the use of the funnel plot of Spiegelhalter (2005a, b) where the performance metric is plotted versus the number of cases. An example of a funnel plot from Symons et al. (2013) is shown in Figure 5. See Mayer et al. (2009) for a discussion of the use of funnel plots in surgical applications.

Because the number of participating hospitals has increased markedly from the inception of NSQIP, the caterpillar plots became unreadable. With this in mind, as well as the above-mentioned shortcoming of the plot, Cohen et al. (2013) reported that NSQIP moved to reporting the odds ratios with confidence intervals, along with the decile in which each hospital lies for each of the adverse events. It is easier to follow progress over time with this information than it is following position on the caterpillar plot.

A variety of methods have been proposed for identifying outlying performers in the comparison of hospitals. Bilimoria et al. (2010) showed that the number of hospitals identified as outlying varies widely depending on the method used. Spiegelhalter (2005) argued that it is important to allow for some overdispersion. It is also important to adjust for the number of hospitals being compared. In this regard, Jones and Spiegelhalter (2008) showed that adjusting the thresholds based on the false discovery rate was better than using the Bonferroni adjustment method. Other discussions of the issues involved in how to classify hospitals as below average, average, or above average were provided by Jones and Spiegelhalter (2011), He at al. (2014), Seaton and Manktelow (2012), Kalbfleisch and Wolfe (2013), Cohen et al. (2013), and Ieva and Paganoni (2015), among others.

One limitation of NSQIP analyses is that the full risk-adjusted outcome reports are provided based on

data aggregated over 6-month intervals with some additional delay for processing. Monitoring on a more continuous basis is possible by running non-risk-adjusted reports on a daily, weekly, or monthly basis. This only allows trending, however, with the assumption that the patient populations are uniform. Providing risk-adjusted charts based on patient-by-patient outcomes, even plotted retrospectively, would likely provide more insight into hospital performance over time and into the effect of improvement initiatives. Cohen et al. (2013) reported, however, that NSQIP is moving toward timelier monitoring.

The Surgical Outcomes Monitoring and Improvement Program of the Hong Kong Hospital Authority is structured similarly to NSQIP. Yuen (2013) reported on their evaluation of 17 public hospitals and their comparison of the observed mortality rates with the expected rates using data from 23,700 operations performed during the period July 2010–June 2012. Elective surgery and emergency surgery were treated separately. Outlying performance was identified using caterpillar plots.

7. NSQIP Case Study

In this section, we report on the surgical quality improvement obtained at the Carilion Clinic in Roanoke, Virginia, with Dr. Sandy L. Fogel, MD, as NSQIP surgeon champion and James Jones, BSN, as surgical clinical reviewer.

Based on the initial NSQIP results in 2007 showing O/E ratios significantly above one, the focus of improvement was in reducing the rate of surgical-site infections and the general surgery mortality. Adverse events tend to be expensive. NSQIP (2014) reported that the cost of a surgical-site infection averages around $27,000, while Dimick et al. (2004) estimated that a case of ventilator-associated pneumonia added about $50,000 to the cost of a surgical admission.

Best practices were used to identify improvement projects. Projects were undertaken to improve each of the following best practices, which were either inconsistently done or not done at all, in order to reduce the rate of surgical-site infections:

- Normothermia (patient warming) throughout the surgical and post-op period
- Post-op glucose control (though EndoTool^{TM})
- Pre-op skin antisepsis at home
- Methicillin-resistant *Staphylococcus aureus* (MRSA) screening and selective decontamination and/or use of Vancomycin for pre-op antibiotic
- Better pre-op glucose control
- Identification and treatment of pre-op infection (especially urinary tract infections)
- Increasing the dose of pre-op antibiotics for obesity
- Redosing antibiotics at 3 hours into procedure
- Transport of post-op patients on oxygen
- Pre-op optimization of respiratory status.

The sequence of implementation of the quality-improvement projects was based on a combination of the expected relative impact on outcomes and the ease of accomplishment, including an assessment of the financial resources needed. The home antisepsis, patient warming, and improved glucose-control projects were implemented first.

It is important to control glucose levels because high blood-sugar levels are associated with higher rates of infection. EndoTool^{TM} is a computerized system for calculating dosing for intravenous insulin. Fogel and Baker (2013) showed use of this computerized system leads to better glucose control than standard paper-based protocols, where insulin doses are calculated using a worksheet. There were seven paper-based protocols at Carilion before the adoption of EndoTool^{TM}, with none used particularly well. The percentage of patients with blood-sugar levels above the high level of 150 milligrams per deciliter (mg/dL) was 31% over a 6-month period before use of EndoTool^{TM} and 16% in the 6-month period afterward. Some patients are insulin-resistant, making it impossible to prevent having some patients with high blood-sugar levels.

With respect to other process-quality variables, the projects led to the following implementation rate changes: home antisepsis, a 20% rate to over 95%; warming, less than 30% to over 95%; redosing in operating room for cases over 3 hours, from 0% to over 75%; and MRSA screening rate of 100%, with MRSA treatment pre-op from 0% to more than 50%.

The dramatic effect of the initiatives in lowering the surgical-site infection rates can be seen in Figure 6. No O/E values from December 2009 onward, marked by open circles, were significantly different from one. Being able to monitor performance over time is a key benefit of participating in NSQIP.

With respect to the general mortality rate, reviews of medical records for prior cases showed that

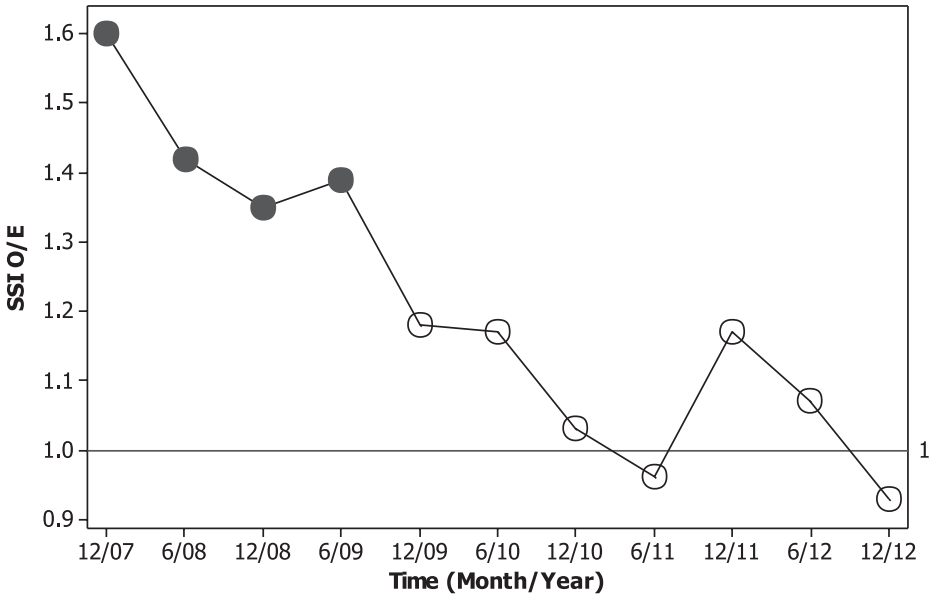

FIGURE 6. Time-Series Plot of Carilion Surgical Site Infection O/E Ratios. Values marked by open circles are not significantly different from one.

some surgical patients were less than medically optimized prior to surgery. This included patients going to surgery with poorly controlled hypertension, diabetes, cardiac disease, etc. Thus, the patient-screening process was moved back from 2–3 days before the planned operation to 2–3 weeks before in order to provide more time for proactive treatment. The effect on the mortality rate of this change and the changes implemented to reduce the rate of surgical-site infections is illustrated in Figure 7. The O/E values from June 2010 onward, marked by open circles, are not significantly different than one, evidence of improvement over earlier performance.

The largest barrier to successful implementation of a given quality-improvement project was surgeon

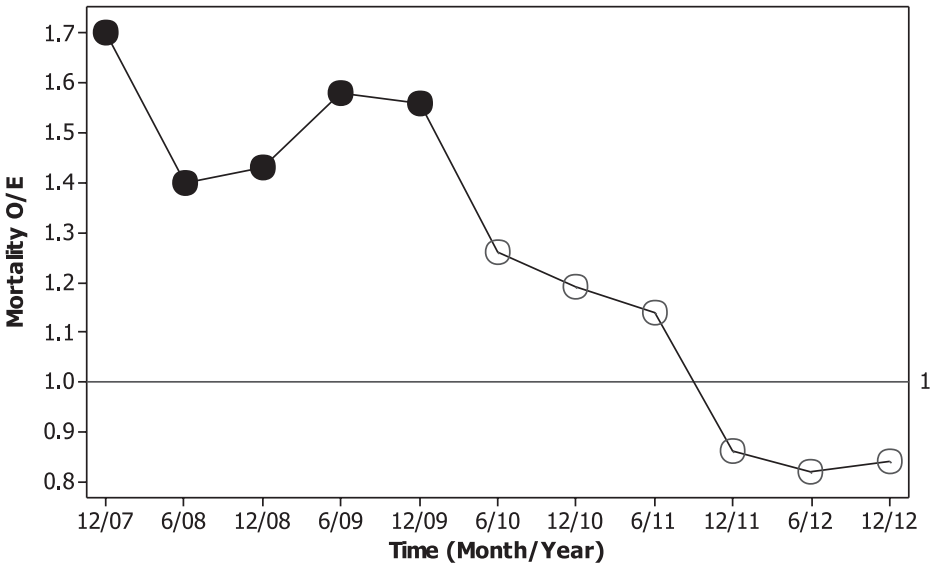

FIGURE 7. Time-Series Plot of Carilion 30-Day Mortality O/E Ratios. Values marked by open circles are not significantly different from one.

habit. It is simply hard to institute a change in behavior. The key was to make the changes as invisible to the surgeon as possible by relying on policies and protocols, with automatic population of the electronic medical record with the appropriate physician orders. Such process changes are difficult, time consuming, and require a great deal of teamwork among individuals and groups to accomplish. Obviously, the surgeons are needed, but so are anesthesiologists, nurses (pre-op, intra-op, and post-op), systems analysts, data analysts, financial representatives, purchasing agents, supply managers, and many others. One of the lessons learned was that it was the processes that needed to be improved, not the performance of any of the surgeons.

There were very significant improvements made in surgical quality that prevented many surgical-site infections and saved many lives. The raw data showed a reduction in mortality from a high of 3.7% to a low of 1.8%. This was a reduction in mortality of approximately 50%. The hospital performs roughly 20,000 surgical procedures a year, which translates into approximately 300 lives saved per year. The need for the improvements was made clear through the NSQIP benchmarking process. The effects of improvement initiatives were then monitored over time using NSQIP reports as process changes were implemented.

8. Some Potential Research Topics

Some topics related to risk-adjusted monitoring that merit further research include the following:

(a) There needs to be studies of alternative methods for risk adjustment, including further study of the use of interaction terms in the logistic regression model. Multiple years of NSQIP data (the Participant Use Data File) are available to NSQIP participants. The 2012 file, for example, contains information on 543,885 cases submitted from 374 participating sites. These data could be used to study the performance of other risk-adjustment approaches. The evolving methodology used by NSQIP was discussed in considerable detail by Cohen et al. (2013).

(b) It is important to study the effect of estimation error on the various monitoring methods, particularly those described in Section 4.6 that incorporate the time until any death within a given time window following surgery. The bootstrap method of Jones and Steiner (2012) and Gandy and Kvaløy (2013) could perhaps be used to control the percentage of the time that the in-control ARL falls below a specified value. This can help to avoid designing charts that result in many false alarms.

(c) It may be possible to build on the work of Yeh et al. (2009) to develop a prospective profile monitoring approach to determine when a risk-adjustment logistic model needs to be updated. Similarly, the change-point approach of Gurevich and Vexler (2005) may be useful in the analysis of the baseline data used to design the surveillance methods, a topic needing more study generally.

(d) There seems to be an opportunity to develop alternatives to the SPRT method for the analysis of learning curves.

(e) The effect of data aggregation on the performance of the various methods needs to be quantified.

(f) Current surveillance methods are based on an assumption of independence of the outcomes. As pointed out by Morton (2003), there could be dependence over time or overdispersion compared with the assumed models. This requires study of current methods under these conditions and the development of new methods. Mousavi and Reynolds (2009) considered the design of a Bernoulli CUSUM chart using a model for dependence over time in the non-risk-adjusted case.

(g) Some applications involve monitoring many process data streams, a topic included in the overview by Woodall and Montgomery (2014). For example, Spiegelhalter et al. (2012) considered the problem of using CUSUM charts to monitor over 200,000 indicators for excess mortality. How to monitor such a large number of data streams most effectively is an area that needs more attention. One must ideally keep the number of false alarms low while maintaining the ability to detect significant outlying performance.

(h) Tang et al. (2015) proposed a method for risk-adjusted monitoring that allows for more than two outcomes. More work is needed in this area. Their method could be designed using the approach of Zhang and Woodall (2015) in order to make the method invariant to the underlying risk distribution.

9. Conclusions

It is clearly important to monitor and improve healthcare quality, which includes surgical quality. We believe that there will be an increasing emphasis on the monitoring and public reporting of risk-adjusted outcome performance metrics, as, for example, in the UK (Bottle and Aylin (2008), Spiegelhalter et al. (2012)). Performance indicators are publicly available for each health trust in the UK. See, for example, Dr. Foster Intelligence (2014).

We strongly encourage hospital administrators to participate in NSQIP or some other collaborative network of hospitals to evaluate their performance results. The business case and the benefits to patients more than justify such participation. For those interested in best surgical practices to improve surgical quality, we also recommend the information provided through the Institute for Healthcare Improvement (www.ihi.org).

In general, we support the monitoring of surgical outcomes on a case-by-case basis with as little aggregation of data over time as possible. Data aggregation can slow the detection of changes in quality and make it more difficult to determine the immediate effects of specific quality-improvement initiatives. The RA-CUSUM chart combined with a VLAD plot is our recommended approach for monitoring on a case-by-case basis with binary data.

Acknowledgments

We thank Dr. Albert Yuen of the Hong Kong Hospital Authority for providing Figure 1. The work of W. H. Woodall was supported by National Science Foundation Grant CMMI-1436365.

References

Albers, W. (2011). "Risk-Adjusted Control Charts for Health Care Monitoring". *International Journal of Mathematics and Mathematical Sciences* Article ID 895273, 16 pages.

Albert, A. A.; Walter, J.A.; Arnrich, B.; Hassanein, W.; Rosendahl, U. P.; Bauer, S.; and Ennker, J. (2004). "On-Line Variable Live-Adjusted Displays with Internal and External Risk-Adjusted Mortalities, A Valuable Method for Benchmarking and Early Detection of Unfavorable Trends in Cardiac Surgery". *European Journal of Cardio-Thoracic Surgery* 25, pp. 312–319.

Alemi, F. and Oliver, D. (2001). "Tutorial on Risk-Adjusted p Charts"". *Quality Management in Health Care* 10, pp. 1–9.

Alemi, F.; Rom, W.; and Eisenstein, E. (1996). "Risk-Adjusted Control Charts for Health Care Assessment". *Annals of Operations Research* 67, pp. 45–60.

Assareh, H. and Mengersen, K. (2014a). "Change Point Estimation in Monitoring Survival Times". *PLoS ONE* 7(3). DOI: 10.1371/journal.pone.0033630.

Assareh, H. and Mengersen, K. (2014b). "Estimation of the Time of a Linear Trend in Monitoring Survival Time". *Health Services and Outcomes Research Methodology* 14(1–2), pp. 15–33.

Axelrod, D. A.; Guidinger, M. K.; Metzger, R. A.; Wiesner, R. H.; Webb, R. L.; and Merion, R. M. (2006). "Transplant Center Quality Assessment Using a Continuously Updatable, Risk-Adjusted Technique (CUSUM)". *American Journal of Transplantation* 6, pp. 313–323.

Axelrod, D. A.; Kalbfleisch, J. D.; Sun, R. J.; Guidinger, M. K.; Biswas, P.; Levine, G. N.; Arrington, C. J.; and Merion, R. M. (2009). "Innovations in the Assessment of Transplant Center Performance: Implications for Quality Improvement". *American Journal of Transplantation* 9(2), pp. 959–969.

Beiles, C. B. and Morton, A. P. (2004). "Cumulative Sum Control Charts for Assessing Performance in Arterial Surgery". *ANZ Journal of Surgery* 74, pp. 146–151.

Benneyan, J. C. (1998a). "Statistical Quality Control Methods in Infection Control and Hospital Epidemiology. Part I: Introduction and Basic Theory". *Infection Control and Hospital Epidemiology* 19, pp. 194–214.

Benneyan, J. C. (1998b). "Statistical Quality Control Methods in Infection Control and Hospital Epidemiology. Part II: Chart Use, Statistical Properties, and Research Issues". *Infection Control and Hospital Epidemiology* 19, pp. 265–283.

Biau, D. J.; Resche-Rigon, M.; Godiris-Petit, G.; Nizard, R. S.; and Porcher, R. (2007). "Quality Control of Surgical and Interventional Procedures: A Review of the CUSUM". *Quality and Safety in Health Care* 16, pp. 203–207.

Bilimoria, K. Y.; Cohen, M. E.; Merkow, R. P.; Wang, X.; Bentrem, D. J.; Ingraham, A. M.; Richards, K.; Hall, B. L.; and Ko, C. Y. (2010). "Comparison of Outlier Identification Methods in Hospital Surgical Quality Improvement Programs". *Journal of Gastrointestinal Surgery* 14, pp. 1600–1607.

Biwas, P. and Kalbfleisch, J. D. (2008). "A Risk-Adjusted CUSUM in Continuous Time Based on the Cox Model". *Statistics in Medicine* 27, pp. 3382–3406.

Blackstone, E. H. (2004). "Monitoring Surgical Performance". *Journal of Thoracic and Cardiovascular Surgery* 128(6), pp. 807–810.

Bolsin, S. and Colson, M. (2000). "The Use of the Cusum Technique in the Assessment of Trainee Competence in New Procedures". *International Journal for Quality in Health Care* 12(5), pp. 433–438.

Bottle, A. and Aylin, P. (2008). "Intelligent Information: A National System for Monitoring Clinical Performance". *Health Services Research* 43, pp. 1–31.

Bruce, J.; Russell, E. M.; Mollison, J.; and Krukowski, Z. H. (2001). "The Measurement and Monitoring of Surgical Adverse Events". *Health Technology Assessment* 5(22).

Chang, T.-C. (2008). "Cumulative Sum Schemes for Surgical Performance Monitoring". *Journal of the Royal Statistical Society, Series A* 171(2), pp. 407–432.

Chen, T.-T.; Chung, K.-P.; Hu, F.-C.; Fan, C. M.; and Yang, M.-C. (2011). "The Use of Statistical Process Control (Risk-Adjusted CUSUM, Risk-Adjusted RSPRT and CRAM with Prediction Limits) for Monitoring the Outcomes of Out-of-Hospital Cardiac Arrest Patients Rescued by the EMS

System". *Journal of Evaluation in Clinical Practice* 17, pp. 71–77.

Clark, D. E.; Hannan, E. L.; and Wu, C. (2010). "Predicting Risk-Adjusted Mortality for Trauma Patients: Logistic Versus Multilevel Logistic Models". *Journal of the American College of Surgeons* 211(2), pp. 224–231.

Clinical Practice Improvement Centre. (2008). *VLADs for Dummies*. Milton, Queensland, Australia: Wiley Publishing Australia Pty Ltd. Available on request from VLAD_Queries@health.qld.gov.au

Cockings, J. G. L.; Cook, D. A.; and Iqbal, R. K. (2006). "Process Monitoring in Intensive Care with the use of Cumulative Expected Minus Observed Mortality and Risk-Adjusted p Charts". *Critical Care* 10, R28. DOI: 10.1186/cc3996.

Cohen, M. E.; Dimick, J. B.; Bilimoria, K. Y.; Ko, C. Y.; Richards, K.; and Hall, B. L. (2009). "Risk Adjustment in the American College of Surgeons National Surgical Quality Improvement Program: A Comparison of Logistic Versus Hierarchical Modeling". *Journal of the American College of Surgeons* 209(6), pp. 687–693.

Cohen, M. E.; Ko, C. Y.; Bilimoria, K. Y.; Zhou, L.; Huffman, K.; Wang, X.; Liu, Y.; Kraemer, K.; Meng, X.; Merkow, R.; Chow, W; Matel, B.; Richards, K.; Hart, A. J.; Dimick, J. B.; and Hall, R. I. (2013). "Optimizing ACS NSQIP Modeling for Evaluation of Surgical Quality and Risk: Patient Risk Adjustment, Procedure Mix Adjustment, Shrinkage Adjustment, and Surgical Focus". *Journal of the American College of Surgery* 217(2), pp. 336–346.

Collett, D.; Sibanda, N.; Pioli, S.; Bradley, A.; and Rudge, C. (2009). "The UK Scheme for Mandatory Continuous Monitoring of Early Transplant Outcome in All Kidney Transplant Centers". *Transplantation* 88, pp. 970–975.

Collins, G. S.; Jibawi, A.; and McCulloch, P. (2011). "Control Charts Methods for Monitoring Surgical Performance: A Case Study from Gastro-Oesophageal Surgery". *European Journal of Surgical Oncology* 37, pp. 473–480.

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, pp. 469–474.

Cook, D. A.; Duke, G.; Hart, G. K.; Pilcher, D.; and Mullany, D. (2008). "Review of the Application of Risk-Adjusted Charts to Analyze Mortality Outcomes in Critical Care." *Critical Care Resuscitation* 10(3), pp. 239–251.

Cook, D. A.; Steiner, S. H.; Cook, R. J.; Farewell, V. T.; and Morton, A. P. (2003). "Monitoring the Evolutionary Process of Quality: Risk-Adjusted Charting to Track Outcomes in Intensive Care". *Critical Care Medicine* 31(6), pp. 1676–1682.

COPPS-CMS White Paper Committee. (2012). *Statistical Issues in Assessing Hospital Performance*. imstat.org/news/2012/03/05/1330972991833.html. Accessed on May 29, 2014.

de Leval, M. R.; Francois, K.; Bull, C.; Brawn, W. B.; and Spiegelhalter, D. J. (1994). "Analysis of a Cluster of Surgical Failures". *Journal of Thoracic and Cardiovascular Surgery* 104, pp. 914–924.

Dimick, J. B.; Chen, S. L.; Taheri, P. A.; Henderson, W. G.; Khuri, S. F.; and Campbell, Jr., D. A. (2004). "Hospital Costs Associated with Surgical Complications: A Report from the Private Sector National Surgical Improvement Program". *Journal of the American College of Surgery* 202(6), pp. 531–537.

Department of Health. (2010). *Equity and Excellence: Liberating the NHS*. London: Department of Health.

Dr. Foster Intelligence (2014). "My Hospital Guide 2013". myhospitalguide.drfosterintelligence.co.uk/#/mortality. Accessed June 2, 2014.

Donabedian, A. (1966). "Evaluating the Quality of Medical Care". *Milbank Memorial Fund Quarterly* 44, pp. 166–206.

Fogel, S. L. and Baker, C. C. (2013). "Effects of Computerized Decision Support Systems on Blood Glucose Regulation in Critically Ill Surgical Patients". *Journal of the American College of Surgeons* 216(4), pp. 828–833.

Gan, F. F.; Lin, L.; and Loke, C. K. (2012). "Risk-Adjusted Cumulative Sum Charting Procedures". In *Frontiers in Statistical Quality Control*, Vol. 10, Lenz, H.-J.; Wilrich, P.-T.; and W. Schmid, W., eds., pp. 207–225. Physica-Verlag.

Gan, F. F. and Tan, T. (2010). "Risk-Adjusted Number-Between Failures Charting Procedures for Monitoring a Patient Care Process for Acute Myocardial Infarctions". *Health Care Management Science* 13, pp. 222–233.

Gandy, A. and Kvaløy, J. T. (2013). "Guaranteed Conditional Performance of Control Charts via Bootstrap Methods". *Scandinavian Journal of Statistics* 40, pp. 647–668.

Gandy, A.; Kvaløy, J. T.; Bottle, A.; and Zhou, F. (2010). "Risk-Adjusted Monitoring of Time to Event". *Biometrika* 97, pp. 375–388.

Gombay, E.; Hussein, A. A.; and Steiner, S. H. (2011). "Monitoring Binary Outcomes Using Risk-Adjusted Charts: A Comparative Study". *Statistics in Medicine* 30, pp. 2815–2826.

Grigg, O. and Farewell, V. (2004a). "An Overview of Risk-Adjusted Charts". *Journal of the Royal Statistical Society, Series A* 167, pp. 523–539.

Grigg, O. and Farewell, V. (2004b). "A Risk-Adjusted Sets Method for Monitoring Adverse Medical Outcomes". *Statistics in Medicine* 23, pp. 1593–1602.

Grigg, O. A.; Farewell, V. T.; and Spiegelhalter, D. J. (2003). "The Use of Risk-Adjusted CUSUM and RSPRT Charts for Monitoring in Medical Contexts". *Statistical Methods in Medical Research* 12, pp. 147–170.

Grigg, O. and Spiegelhalter, D. J. (2007). "Simple Risk-Adjusted Exponentially Weighted Moving Average". *Journal of the American Statistical Association* 102, pp. 140–152.

Grunkemeier, G. L.; Jin, R.; and Wu, Y. (2009). "Cumulative Sum Curves and Their Prediction Limits". *Annals of Thoracic Surgery* 87, pp. 361–364.

Grunkemeier, G. L.; Wu, Y. X.; and Furnary, A. P. (2003). "Cumulative Sum Techniques for Assessing Surgical Results". *Annals of Thoracic Surgery* 76, pp. 663–667.

Gurevich, G. and Vexler, A. (2005). "Change Point Problems in the Model of Logistic Regression". *Journal of Statistical Planning and Inference* 131, pp. 313–331.

Gustafson, T. L. (2000). "Practical Risk-Adjusted Quality Control Charts for Infection Control". *American Journal of Infection Control* 28(6), pp. 406–414.

Hall, B. L.; Hamilton, B. H.; Richards, K.; Bilimoria, K. Y.; Cohen, M. E.; and Ko, C.Y. (2009). "Does Surgical Quality Improve in the American College of Surgeons National Surgical Quality Improvement Program: An Evaluation of All Participating Hospitals". *Annals of Surgery* 205(3), pp. 363–376.

Harris, J. R.; Forbes, T. L.; Steiner, S. H.; Lawlor, K.; Derose, G. and Harris, K. A. (2005). "Risk-Adjusted Analysis of Early Mortality After Ruptured Abdominal Aortic Aneurysm Repair". *Journal of Vascular Surgery* 42, pp. 387–391.

Hart, M. K.; Lee, K. Y.; Hart, R. F.; and Robertson, J. W. (2003). "Application of Attribute Control Charts to Risk-Adjusted Data for Monitoring and Improving Health Care Performance". *Quality Management in Health Care* 12(1), pp. 5–19.

Hart, M. K.; Robertson, J. W.; Hart, R. F.; and Lee, K. Y. (2004). "Application of Variables Control Charts to Risk-Adjusted Time-Ordered Healthcare Data". *Quality Management in Health Care* 13(2), pp. 99–119.

He, Y.; Selck, F.; and Sharon-Lise, T. N. (2014). "On the Accuracy of Classifying Hospitals on Their Performance Measures". *Statistics in Medicine* 33, pp. 1081–1103.

Ieva, F. and Paganori, A. M. (2015). "Detecting and Visualizing Outliers in Provider Profiling via Funnel Plots and Mixed Effect Models". *Health Care Management Science*, to appear. DOI 10.1007/s10729-013-9264-9.

Iezzoni, L. (2012). *Risk Adjustment for Measuring Health Care Outcomes*, 4th edition. Chicago, IL: Health Administration Press.

Institute for Healthcare Improvement (2014). "Surgical Site Infection". www.ihi.org/Topics/SSI/Pages/default.aspx. Accessed June 1, 2014.

Jones, H. E. and Spiegelhalter, D. J. (2008). "Use of False Discovery Rate When Comparing Multiple Healthcare Providers". *Journal of Clinical Epidemiology* 61(3), pp. 232–240.

Jones, H. E. and Spiegelhalter, D. J. (2011). "The Identification of 'Unusual' Health-Care Providers from a Hierarchical Model". *The American Statistician* 65(3), pp. 154–163.

Jones, M. A. 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(2), pp. 176–181.

Jones-Farmer, L. A.; Woodall, W. H.; Steiner, S. H.; and Champ, C. W. (2014). "An Overview of Phase I Analysis for Process Improvement and Monitoring". *Journal of Quality Technology* 46(3), pp. 265–280.

Kalbfleisch, J. D. and Wolfe, R. A. (2013). "On Monitoring Outcomes of Medical Providers". *Statistics in Biosciences* 5(2), pp. 286–302.

Ko, C. Y. (2009). "Measuring and Improving Surgical Quality". *Patient Safety and Quality Healthcare* 6(6), pp. 36–41.

Leandro, G.; Rolando, N.; Gallus, G.; Rolles, K.; and Burroughs, A. K. (2005). "Monitoring Surgical and Medical Outcomes: The Bernoulli Cumulative SUM Chart. A Novel Application to Assess Clinical Interventions". *Postgraduate Medical Journal* 81, pp. 647–652.

Loke, C. K. and Gan, F. F. (2012). "Joint Monitoring Scheme for Clinical Failures and Predisposed Risks". *Quality Technology and Quantitative Management* 9(1), pp. 3–21.

Lovegrove, J.; Sherlaw-Johnson, C.; Valencia, O.; Treasure, T.; and Gallivan, S. (1999). "Monitoring the Performance of Cardiac Surgeons". *Journal of the Operational Research Society* 50, pp. 684–689.

Lovegrove, J.; Valencia, O.; Treasure, T.; Sherlaw-Johnson, C.; and Gallivan, S. (1997). "Monitoring the Results of Cardiac Surgery by Variable Life-Adjusted Display". *Lancet* 18, pp. 1128–1130.

Maggard-Gibbons, M. (2013). "Use of Report Cards and Outcome Measurements to Improve Safety of Surgical Care: American College of Surgeons National Quality Improvement Program". Chapter 14 in *Making Healthcare Safer. II: An Updated Critical Analysis of the Evidence for Patient Safety Practices, Evidence Reports/Technology Assessments*, No. 211, Report No. 13-E001-EF, pp. 140–157. Rockville, MD: Agency for Healthcare Research and Quality.

Mayer, E. K.; Bottle, A.; Rao, C.; Darsi, A. W.; and Thanos, A. (2009). "Funnel Plots and Their Emerging Application in Surgery". *Annals of Surgery* 249(3), pp. 376–383.

Megahed, F. M.; Kensler, J. L. K.; Bedair, K.; and Woodall, W. H. (2011). "A Note on the ARL of Two-Sided Bernoulli-Based CUSUM Control Charts". *Journal of Quality Technology* 43(1), pp. 43–49.

Moore, R.; Nutley, M.; Cina, C. S.; Motamedi, M.; Faris, P.; and Abuznadah, W. (2007). "Improved Survival After Introduction of an Emergency Endovascular Therapy Protocol for Ruptured Abdominal Aortic Aneurysms". *Journal of Vascular Surgery* 4, pp. 443–450.

Morton, A. P. (2003). "The Use of Statistical Process Control Methods in Monitoring Clinical Performance—Letter to the Editor". *International Journal for Quality in Health Care* 15(4), pp. 361–362.

Morton, A. P.; Clements, A. C. A.; Doidge, S. R.; Stackelroth, J.; Curtis, M.; and Whitby, M. (2008). "Surveillance of Healthcare-Acquired Infections in Queensland, Australia: Data and Lessons Learned in the First 5 Years". *Infection Control and Hospital Epidemiology* 29(8), pp. 695–701.

Mousavi, S. and Reynolds, M. R., Jr. (2009). "A CUSUM Chart for Monitoring a Proportion with Autocorrelated Binary Observations". *Journal of Quality Technology* 41(4), pp. 401–414.

Moustakides, G. V. (1986). "Optimal Stopping Times for Detecting Changes in Distribution". *The Annals of Statistics* 14, pp. 1379–1387.

Nashef, S. A. M.; Roques, F.; Sharples, L. D.; Nilsson, J.; Smith, C.; Goldstone, A. R.; and Lockowandt, U. (2012). "EuroSCORE II". *European Journal of Cardio-Thoracic Surgery* 41(4), pp. 734–745.

National Surgical Quality Improvement Program (2014) site.acsnsqip.org/about/business-case/. Accessed on May 29, 2014.

Neuberger, J.; Madden, S.; and Collett, D. (2010). "Review of Methods for Measuring and Comparing Center Performance After Organ Transplantation". *Liver Transplantation* 16, pp. 1119–1128.

New York State Department of Health (2001). *Coronary Artery Bypass Surgery in New York State 1996–1998*. www.health.ny.gov/statistics/diseases/cardiovascular/heart_disease/docs/1996-1998_adult_cardiac_surgery.pdf. Accessed on May 29, 2014.

Novick, R. J.; Fox, S. A.; Stitt, L. W.; Forbes, T. L.; and Steiner, S. H. (2006). "Direct Comparison of Risk-Adjusted and Non-Risk-Adjusted CUSUM Analyses of Coronary Artery Bypass Surgery Outcomes". *Journal of Thoracic and Cardiovascular Surgery* 132, pp. 386–391.

Novick, R. J.; Fox, S. A.; Stitt, L. W.; Swinamer, S. A.; Lehnhardt, K. R.; Rayman, R.; and Boyd, W. D. (2001). "Cumulative Sum Failure Analysis of a Policy Change from On-Pump to Off-Pump Coronary Artery Bypass Grafting". *The Annals of Thoracic Surgery* 72(3), pp. S1016–S1021.

Novick, R. J. and Stitt, L. W. (1999). "The Learning Curve of an Academic Cardiac Surgeon: Use of the CUSUM Method". *Journal of Cardiac Surgery* 14(5), pp. 312–320.

Noyez, L. (2009). "Control Charts, Cusum Techniques and Funnel Plots. A Review of Methods for Monitoring Performance in Healthcare". *Interactive Cardiovascular and Thoracic Surgery* 9, pp. 494–499.

Page, E. S. (1954). "Continuous Inspection Schemes". *Biometrika* 41, pp. 100–114.

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(Supplement 1), pp. 1–12.

Paynabar, K.; Jin, J. H.; and Yeh, A. B. (2012). "Phase I Risk-Adjusted Control Charts for Monitoring Surgical Performance by Considering Categorical Covariates". *Journal of Quality Technology* 44(1), pp. 39–53.

Phinikettos, I. and Gandy, A. (2014). "An Omnibus CUSUM Chart for Monitoring Time to Event Data". *Lifetime Data Analysis* 20, pp. 481–494.

Poloniecki, J.; Sismanidis, C.; Bland, M.; and Jones, P. (2004). "Retrospective Cohort Study of False Alarms Associated with a Series of Heart Operations: The Case for Hospital Mortality Monitoring Groups". *British Medical Journal* 328, pp. 375–378.

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, pp. 1697–1700.

Porter, M. E. and Teisberg, E. O. (2007). "How Physicians Can Change the Future of Health Care". *Journal of the American Medical Association* 297(10), pp. 1103–1111.

Reynolds, M. R., Jr. and Stoumbos, Z. G. (1999). "A CUSUM Chart for Monitoring a Proportion when Inspecting Continuously". *Journal of Quality Technology* 31, pp. 87–108.

Rogers, C. A.; Ganesh, J. S.; Nicholas, R.; Banner, N. R.; and Bonser, R. S. (2005). "Cumulative Risk Adjusted Monitoring of 30-Day Mortality After Cardiothoracic Transplantation: UK Experience". *European Journal of Cardio-Thoracic Surgery* 27, pp. 1022–1029

Rogers, C. A.; Reeves, B. C.; Caputo, M.; Ganesh, J. S.; Bonser, R. S.; and Angelini, G. D. (2004). "Control Chart Methods for Monitoring Cardiac Surgical Performance and Their Interpretation". *Journal of Thoracic and Cardiovascular Surgery* 128, pp. 811–819.

Schuh, A.; Woodall, W. H.; and Camelio, J. A. (2013). "The Effect of Aggregating Data When Monitoring a Poisson Process". *Journal of Quality Technology* 45(3), pp. 260–272.

Seaton, S. E. and Manktelow, B. N. (2012). "The Probability of Being Identified as an Outlier with Commonly Used Funnel Plot Control Limits for the Standardized Mortality Ratio". *BMC Medical Research Methodology* 12, p. 98.

Sego, L H.; Reynolds, M. R., Jr.; and Woodall, W. H. (2009). "Risk-Adjusted Monitoring of Survival Times". *Statistics in Medicine* 28, pp. 1386–1401.

Sego, L. H.; Woodall, W. H.; and Reynolds, M. R., Jr. (2008). "A Comparison of Surveillance Methods for Small Incidence Rates". *Statistics in Medicine* 27(8), pp. 1225–1247.

Shen, X.; Tsung, F.; Zou, C.; and Jiang, W. (2013). "Monitoring Poisson Count Data with Probability Control Limits When Sample Sizes Are Time-varying". *Naval Research Logistics* 60(8), pp. 625–636.

Sherlaw-Johnson, C. (2005). "A Method for Detecting Runs of Good and Bad Clinical Outcomes on Variable Life-Adjusted Display (VLAD) Charts". *Health Care Management Science* 8, pp. 61–65.

Sherlaw-Johnson, C.; Lovegrove, J.; Treasure, T.; and Gallivan, S. (2000). "Likely Variations in Perioperative Mortality Associated with Cardiac Surgery: When Does High Mortality Reflect Bad Practice?" *Heart* 84, pp. 79–82.

Sherlaw-Johnson, C.; Morton, A.; Robison, M. B.; and Hall, A. (2005). "Real-Time Monitoring of Coronary Care Mortality: A Comparison and Combination of Two Monitoring Tools". *International Journal of Cardiology* 100, pp. 301–307.

Sherlaw-Johnson, C.; Wilson, P.; and Gallivan, S. (2007). "The Development and Use of Tools for Monitoring the Occurrence of Surgical Wound Infections". *Journal of the Operational Research Society* 58, pp. 228–234.

Sismanidis, C.; Bland, M.; and Poloniecki, J. (2003). "Properties of the Cumulative Risk-Adjusted Mortality (CRAM) Chart, Including the Number of Deaths Before a Doubling of the Death Rate Is Detected". *Medical Decision Making* 23(3), pp. 242–251.

Snyder, J. J.; Salkowski, N.; Zaun, D.; Leppke, S. N.; Leighton, T.; Israni, A. K.; and Kasiske, B. L. (2014). "New Quality Monitoring Tools Provided by the Scientific Registry of Transplant Recipients: CUSUM". *American Journal of Transplantation* 14, pp. 515–523.

Spiegelhalter, D. J. (2005a). "Funnel Plots for Comparing Institutional Performance". *Statistics in Medicine* 24, pp. 1185–1202.

Spiegelhalter, D. J. (2005b). "Handling Over-Dispersion of Performance Indicators". *Quality and Safety in Healthcare* 14, pp. 347–351.

Spiegelhalter, D.; Grigg, O.; Kinsman, R.; and Treasure, T. (2003). "Risk-Adjusted Sequential Probability Ratio Tests: Applications to Bristol, Shipman and Adult Cardiac Surgery". *International Journal for Quality in Health Care* 15, pp. 7–13.

Spiegelhalter, D.; Sherlaw-Johnson, C.; Bardsley, M.; Blunt, I.; Wood, C.; and Grigg, O. (2012). "Statistical Methods for Healthcare Regulation: Rating, Screening and Surveillance (with Discussion)". *Journal of the Royal Statistical Society, Series A* 175(1), pp. 1–47.

Steiner, S. H. (2014). "Risk-Adjusted Monitoring of Outcomes in Health Care". Chapter 14 in *Statistics in Action: A Canadian Outlook*, Lawless, J. F., ed., pp. 245–264. Chapman and Hall/CRC.

Steiner, S. H.; Cook, R. J.; and Farewell, V. T. (1999). "Monitoring Paired Binary Surgical Outcomes Using Cumulative Sum Charts". *Statistics in Medicine* 18, pp. 69–86.

Steiner, S. H.; Cook, R. J.; and Farewell, V. T. (2001). "Risk-Adjusted Monitoring of Binary Surgical Outcomes". *Medical Decision Making* 21(3), pp. 163–169.

Steiner, S. H.; Cook, R. J.; Farewell, V. T.; and Treasure, T. (2000). "Monitoring Surgical Performance Using Risk-Adjusted Cumulative Sum Charts". *Biostatistics* 1, pp. 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, pp. 444–454.

Steiner, S. H. and MacKay, R. J. (2014). "Monitoring Risk-Adjusted Medical Outcomes Allowing for Changes over Time". *Biostatistics* 15(4), pp. 665–676.

Sun, R. J. and Kalbfleisch, J. D. (2013). "A Risk-Adjusted O-E CUSUM with Monitoring Bands for Monitoring Medical Outcomes". *Biometrics* 69, pp. 62–69.

Symons, N. R. A.; Moorthy, K.; Almoudaris, A. M.; Bottle, A.; Aylin, P.; Vincent, C. A.; and Faiz, O. D. (2013). "Mortality in High-Risk Emergency General Surgical Admissions". *British Journal of Surgery* 100, pp. 1318–1325.

Szarka, III, J. L. and Woodall, W. H. (2011). "A Review and Perspective on Surveillance of Bernoulli Processes". *Quality and Reliability Engineering International* 27, pp. 735–752.

Tang, X.; Gan, F. F.; and Zhang, L. (2015). "Risk-Adjusted Cumulative Sum Charting Procedure Based on Multi-Responses". *Journal of American Statistical Association*, to appear.

Tian, W.; Sun, H.; Zhang, X.; and Woodall, W. H. (2015). "The Impact of Varying Patient Populations on the In-Control Performance of the Risk-Adjusted Bernoulli CUSUM Chart". *International Journal for Quality in Health Care* 27(1), pp. 31–36.

Treasure, T.; Gallivan, S.; and Sherlaw-Johnson, C. (2004). "Monitoring Cardiac Surgical Performance: A Commentary". *Journal of Thoracic and Cardiovascular Surgery* 128, pp. 823–825.

U.S. Department of Health and Human Services. (2013). "National Action Plan to Prevent Health Care-Associated Infections: Roadmap to Elimination". www.health.gov/hai/prevent_hai.asp. Accessed June 6, 2014.

Wald, A. (1947). *Sequential Analysis*. New York, NY: Wiley.

Webster, R. A. and Pettitt, A. N. (2007). "Stability of Approximations of Average Run Length of Risk-Adjusted CUSUM Schemes Using the Markov Approach: Comparing Two Methods of Calculating Transition Probabilities". *Communications in Statistics—Simulation and Computation* 36, pp. 471–482.

Williams, S. M.; Parry, B. R.; and Schlup, M. M. T. (1992). "Quality Control: An Application of the CUSUM". *British Medical Journal* 304, pp. 1359–1361.

Winkel, P. and Zhang, N. F. (2007). *Statistical Development of Quality in Medicine*. Hoboken, NJ: John Wiley & Sons, Inc.

Winkel, P. and Zhang, N. F. (2012). "Statistical Process Control in Clinical Medicine", Chapter 15 in *Statistical Methods in Healthcare*, Faltin, F. W.; Kenett, R.; and. Ruggeri, F., eds., pp. 309–334. John Wiley & Sons, Inc.

Woodall, W. H. (2006). "Use of Control Charts in Health-Care and Public-Health Surveillance (with Discussion)". *Journal of Quality Technology* 38(2), pp. 89–104.

Woodall, W. H.; Adams, B. M.; and Benneyan, J. C. (2012). "The Use of Control Charts in Healthcare", Chapter 12 in *Statistical Methods in Healthcare*, Faltin, F. W.; Kenett, R.; and. Ruggeri, F., eds., pp. 253–267. John Wiley & Sons, Inc.

Woodall, W. H. and Montgomery, D. C. (2014). "Some Current Directions in the Theory and Application of Statistical Process Monitoring". *Journal of Quality Technology* 46(1), pp. 78–94.

Yap, C.-H.; Colson, M. E.; and Watters, D. A. (2007). "Cumulative Sum Techniques for Surgeons: A Brief Review". *ANZ Journal of Surgery* 77, pp. 583–586.

Yeh, A. B.; Huwang, L.; and Li, Y.-M. (2009). "Profile Monitoring for a Binary Response". *IIE Transactions* 41(11), pp. 931–941.

Yuen, W.-C. (2013). "Applying Variable Life Adjusted Display in Monitoring Surgical Outcomes". Paper presented at the 2013 Hong Kong Hospital Authority Convention. www.ha.org.hk/haconvention/hac2013/proceedings/downloads/MC1.1.pdf. Accessed on June 1, 2014.

Zeng, L. and Zhou, S. (2011). "A Bayesian Approach to Risk-Adjusted Outcome Monitoring in Healthcare". *Statistics in Medicine* 30, pp. 3431–3446.

Zhang, X. and Woodall, W. H. (2015). "Dynamic Control Limits for the Risk-Adjusted Bernoulli CUSUM Chart". *Statistics in Medicine*, to appear.