J. R. Statist. Soc. A (2004)
167, *Part* 3, *pp.* 523–539

An overview of risk-adjusted charts

O. Grigg and V. Farewell

Medical Research Council Biostatistics Unit, Cambridge, UK

[Received February 2003. Revised October 2003]

Summary. The paper provides an overview of risk-adjusted charts, with examples based on two data sets: the first consisting of outcomes following cardiac surgery and patient factors contributing to the Parsonnet score; the second being age–sex-adjusted death-rates per year under a single general practitioner. Charts presented include the cumulative sum (CUSUM), resetting sequential probability ratio test, the sets method and Shewhart chart. Comparisons between the charts are made. Estimation of the process parameter and two-sided charts are also discussed. The CUSUM is found to be the least efficient, under the average run length (ARL) criterion, of the resetting sequential probability ratio test class of charts, but the ARL criterion is thought not to be sensible for comparisons within that class. An empirical comparison of the sets method and CUSUM, for binary data, shows that the sets method is more efficient when the in-control ARL is small and more efficient for a slightly larger range of in-control ARLs when the change in parameter being tested for is larger. The Shewart p-chart is found to be less efficient than the CUSUM even when the change in parameter being tested for is large.

Keywords: Average run length; Cumulative sum; Performance; Resetting sequential probability ratio test; Risk-adjusted control charts; Sets method; Shewhart

1. Introduction

Quality control originated in the industrial context, where quick detection of problems is essential for efficiency. Control charts, such as the Shewhart chart and cumulative sum (CUSUM) chart, are a primary statistical tool of the methodology and have been used to monitor automated processes since the 1920s. Recently, it has been suggested that such monitoring schemes could be used to monitor the performance of clinical practitioners, such as surgeons and general practitioners (DeLeval *et al.*, 1994; Lovegrove *et al.*, 1997, 1999; Poloniecki *et al.*, 1998; Steiner *et al.*, 2000; Spiegelhalter *et al.*, 2003).

Unlike in industrial processes, where the 'subjects' (raw material) may be relatively homogeneous in nature, for medical applications the subjects (patients) will often vary greatly in terms of base-line risk. If the heterogeneity of the base-line patient risk is not taken into account when monitoring the surgeon's performance, then the additional variability in outcome due to that heterogeneity may mask the effect of the underlying performance of the surgeon and cause the chart either to produce false alarms or not to respond quickly to changes in performance. Such risk adjustment (adjustment for patient case mix) has been implemented for cumulative observed–expected (O–E) plots by Lovegrove *et al.* (1997, 1999) and Poloniecki *et al.* (1998), CUSUM charts (Steiner *et al.*, 2000; Spiegelhalter *et al.*, 2003), resetting sequential probability ratio test (RSPRT) charts (Spiegelhalter *et al.*, 2003), Shewhart charts (Cook *et al.*, 2003) and the sets method (Grigg and Farewell, 2004).

Address for correspondence: Olivia Grigg, Medical Research Council Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, UK.
E-mail: olivia.grigg@mrc-bsu.cam.ac.uk

© 2004 Royal Statistical Society 0964–1998/04/167523

1.1. Example data

Two example data sets are used throughout. The first is based on data from a centre for cardiac surgery, collected over the period 1992–1998. The data consist of 30-day mortality following surgery, age, sex, type of operation performed, diabetes status and Parsonnet score calculated from the first four variables plus others (see Parsonnet *et al.* (1989)). Monitoring the data by using risk-adjusted CUSUMs is discussed in detail by Steiner *et al.* (2000).

Here, the data that relate to patients of just one of the surgeons are discussed. Patient risk is taken into account by using a logistic regression model relating the probability of surgical failure to patient factors, i.e. a model is used to predict a patient's risk of dying within 30 days of their operation, given their characteristics. As in Steiner *et al.* (2000), a model fitted to the data from 1992–1993 is used to set up charts to monitor the data (retrospectively) from 1994–1998. The average failure rate, from the start of 1994, is assumed to be 0.066. The choice of the model by use of backward elimination resulted in the Parsonnet score being the only factor included. Since the Parsonnet score is a measure that is based on the other factors, and so is highly correlated with them, this is unsurprising.

For this example, since the data are binary, the natural parameter on which to base monitoring is the odds of failure. An acceptable level of performance is taken to be that reflected by the logistic regression model. Departures from this performance level are defined, relative to the model, by a common increase, or decrease, in the odds of death for all patients. Because the early detection of a problem is crucial in this instance, monitoring patient by patient is illustrated, except the Shewhart p-chart which updates every 79 patients (approximately 6 months of surgery for a typical surgeon, on the basis of the training data).

The second example is based on the number of deaths per year, in the period 1987–1998, of the patients of a single general practitioner, Harold Shipman. A public inquiry concluded that Shipman killed at least 215 of his patients over 23 years, a rate of over nine patients a year (Shipman Inquiry, 2002).

Monitoring charts are based on the assumption that the number of deaths per year is Poisson distributed. The acceptable level of 'risk' for a specific patient of type t (male or female; aged under 45, 45–64, 65–74, 75–84 or over 84 years) is taken to be the England and Wales average rate of deaths per year for that type of patient, as given by Baker (2001). This is multiplied by the number of patients of type t in Shipman's practice to give an expected count for the assumed Poisson distribution. A chart that is unadjusted for risk would assume that the risk for each patient is equal to a weighted average over all types of patient of the average rates for England and Wales. It is assumed that the acceptable rates remain the same over the period 1987–1998. In principle, though, we could forecast the trend in rates and allow the expected rate to change over time.

For example 2, since the data are counts, the natural parameter for monitoring is the rate per year (or risk). The null rates are taken to be the rates for England and Wales adjusted for the age–sex distribution of patients under Shipman's care. Departures from this level are defined to be an increase, or decrease, in the risk of death for all patients. Since the rates are assumed to be Poisson distributed, they can be combined into a single rate.

2. Departure from the model: observed–expected plot

Risk-adjusted O–E plots were developed by both Poloniecki *et al.* (1998) (who named them cumulative-risk-adjusted mortality charts) and Lovegrove *et al.* (1997) (who referred to them as variable-life-adjusted display plots).

Fig. 1 compares an unadjusted O–E plot with a risk-adjusted O–E plot. This chart is based on example data set 1, where the patients' outcome is a binary indicator Y_t of whether patient t

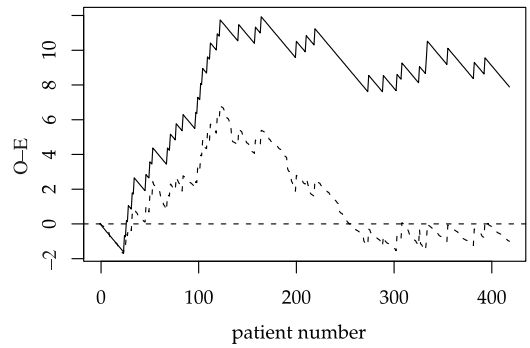

Fig. 1. O–E plot of surgical outcomes (1994–1998) for a cardiac surgeon: ———, unadjusted; -------, risk adjusted

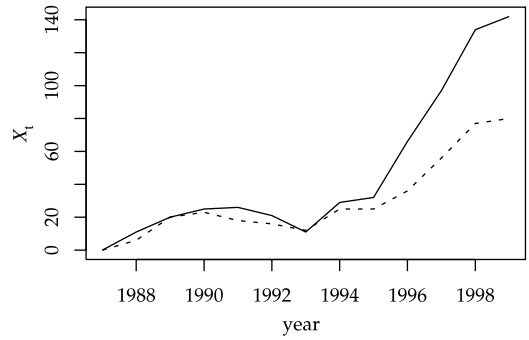

Fig. 2. O–E plot of deaths under Harold Shipman (1987–1998), where the expected value is based on the average for England and Wales: ———, all patients; -------, females aged 75 years and over

survived 30 days after cardiac surgery. The unadjusted plot assumes that the risk of death for all patients is the same value p (the risk averaged over the type of patient) and plots the cumulative sum, over time, of $y_t - p$. In contrast, the adjusted O–E plot presents the cumulative sum, over time, of $y_t - p_t$, where p_t is the estimated risk of death from the logistic regression model.

The plot that is unadjusted for risk suggests that the surgeon is performing much more poorly than the risk-adjusted version suggests. This is because many of these operations were on high risk patients. Only by taking this into account can we accurately assess the surgeon's performance.

Fig. 2 shows a risk-adjusted O–E plot for the example 2 data over the period 1987–1998. Also shown is a comparable plot for females aged 75 years and over. Here risk adjustment means a comparison of the observed rates under Shipman *for each category of patient* with the corresponding expected rate based on rates for England and Wales. In this plot this corresponds to a simple adjustment which is made by calculating, for each year, the age and sex adjusted rate for the patient mix in Harold Shipman's practice and then averaging over the years to give a single rate which is used throughout the chart. For the all-patient chart this is 35 deaths per year and for the females over 75 years chart it is 12 deaths per year. The rapid rise in the death-rate in the latter years is clear. The increase in the overall death-rate appears to be mainly attributable to the rise in rate for females who were over 75 years old; in 1989 and 1993, in particular, the calculated overall excess is wholly due to an excess in deaths of females over 75 years of age.

The cumulative O–E statistic represents an intuitively useful way to represent performance over time. However, this type of plot is not the most natural from which to determine if and

when an alarm should be raised. The CUSUM and RSPRT charts, which are similar to the O–E plot, are designed with this purpose in mind.

3. Resetting sequential probability ratio test and cumulative sum charts

The RSPRT and CUSUM charts are both derived from the Wald sequential probability ratio test (SPRT) (Wald, 1945). The SPRT is a sequential test of a hypothesis H_0 *versus* an alternative H_1. The test statistic is the log-likelihood ratio X_t in favour of H_1 of the cumulative data up to and including time t. The value of X_t can be expressed as

$$X_t = X_{t-1} + L_t, \qquad t = 1, 2, 3, \ldots \tag{1}$$

where $X_0 = 0$ and L_t is the log-likelihood ratio for the single data point at time t.

The SPRT terminates in favour of hypothesis H_0 if the lower boundary a is crossed with approximate type I error rate α and in favour of H_1 if the upper boundary b is crossed with approximate type II error rate β, where

$$\begin{aligned} a &= \log\left(\frac{\beta}{1-\alpha}\right), \\ b &= \log\left(\frac{1-\beta}{\alpha}\right). \end{aligned} \tag{2}$$

If the data contain risk information, this can be taken into account in the test through the likelihood. For example 1, the risk model relating 30-day mortality to Parsonnet score V_t was taken to be

$$\text{logit}(p_r) = -3.67 + 0.077 V_r, \qquad r = 1, 2, \ldots, m, \tag{3}$$

where p_r is the probability of a patient of type r failing within 30 days of surgery. Consider the hypotheses to be H_0: $p_{r0} = p_r$ and H_1: $p_{r1} = R p_r / \{1 + (R-1) p_r\}$, $r = 1, 2, \ldots, m$. If the data are assumed to be Bernoulli distributed, then the log-likelihood ratio for the SPRT would be taken to be

$$L_t = \log\left\{\frac{p_{r1}^{y_t}(1-p_{r1})^{1-y_t}}{p_{r0}^{y_t}(1-p_{r0})^{1-y_t}}\right\} \tag{4}$$

where y_t is the outcome for the tth patient.

For example 2, if we let $\lambda = \Sigma_{s=1}^{10} \lambda_s$ be the combined death-rate for all types of patient (where λ_s is the rate for a patient of type s, $s = 1, 2, \ldots, 10$) and define the hypotheses to be H_0: $\lambda_0 = \lambda$ and H_1: $\lambda_1 = R\lambda$, then (under the assumption that the data are Poisson distributed), the log-likelihood ratio for the SPRT would be

$$L_t = \log\left\{\frac{\lambda_1^{y_t}\ \exp(-\lambda_1)}{\lambda_0^{y_t}\ \exp(-\lambda_0)}\right\}. \tag{5}$$

3.1. Cumulative sum chart

The CUSUM (strictly, the *tabular* CUSUM) was developed by Page (1954). As with the Wald SPRT, the cumulative log-likelihood ratio is plotted, but in this case H_0 is viewed as a null hypothesis. Because the chart's intended purpose is on-going monitoring and not a single significance test, acceptance of the null hypothesis makes little sense. The chart is prevented from

crossing the lower boundary and accepting hypothesis H_0 by replacing the lower absorbing barrier at a with a holding barrier at 0.

For the CUSUM, the cumulative log-likelihood ratio for data up to and including time t can be written as

$$X_t = \max(0, X_{t-1} + L_t), \qquad t = 1, 2, 3, \ldots, \tag{6}$$

where, as for the SPRT, $X_0 = 0$ and L_t is the log-likelihood ratio for the single data point at time t. The chart is said to 'signal' when $X_t > h$, where h defines an upper boundary for the plot. At this point, it is expected that monitoring will stop and remedial action will be taken.

The performance of an SPRT is determined by its nominal error rates α and β, whereas the efficiency of a CUSUM chart is quantified in terms of the length of time before an alarm, false or true, is raised. The average run length (ARL) to detection of an alarm is a convenient and common criterion that is used to assess a CUSUM's performance. The ARL to detection when the process is in state H_0 is termed the in-control ARL and this is analogous to the type I error rate of an SPRT. The out-of-control ARL is analogous to the type II error rate of an SPRT. Typically, the in-control ARL is fixed by setting the boundary h and then the out-of-control ARL is measured for a chart with that same boundary.

3.2. Resetting sequential probability ratio test chart

A more flexible class of charts, which includes the CUSUM as a special case, is the RSPRT chart, suggested by Spiegelhalter *et al.* (2003) and discussed in detail by Grigg *et al.* (2003). These, like the CUSUM, are also based on the SPRT, but rather than having a lower holding barrier at 0 have a lower highly elastic (or resetting) barrier at a, i.e., when the lower boundary a is reached, the chart resets to 0 and monitoring continues. So, where the CUSUM can be viewed as a sequence of SPRTs with lower boundary at 0 and upper boundary at h, an RSPRT can be viewed as a sequence of SPRTs with lower boundary at a and upper boundary at b. Hence the CUSUM is an RSPRT with $a = 0$.

Note that the barriers (a, b) can be defined by parameters (α^*, β^*) through equations (2), replacing α and β with α^* and β^* respectively. The pair (α^*, β^*) are simply parameters and are chosen for convenience, and, unlike for the non-resetting SPRT, have no relationship to the type I and II error rates α and β of the chart. Because the chart resets until the upper boundary is crossed, the actual type I and II error rates for an RSPRT are in fact 1 and 0 respectively.

3.2.1. Optimizing α^* and β^*, the parameters of the resetting sequential probability ratio test

There is an infinite number of pairs (α^*, β^*) (defined by equations (2) where $a \approx 0$ and $b \approx h$) that give the same in-control ARL as a CUSUM with control limit h, but only a small number of those (giving boundaries close to $(0, h)$) have the same out-of-control ARL. Note, however, that there are RSPRT charts that have a smaller out-of-control ARL than the CUSUM with the same in-control ARL. Essentially, a small out-of-control ARL can be achieved by having a relatively low upper boundary b which is smaller in absolute value than the lower boundary a. This characteristic is achieved when β^* is chosen to be very small compared with α^*.

Considering example data 1, Fig. 3 shows, for risk-adjusted RSPRTs that are designed to detect a doubling of the odds of 30-day mortality, how the out-of-control ARL varies for various choices of (α^*, β^*) which give the same in-control ARL. The in-control ARL for all pairs is approximately 6700 patients, which is equivalent to roughly 6 years of surgery. The out-of-control ARL, given alongside selected points (α^*, β^*), is seen to decrease with increasing α^* and decreasing β^*.

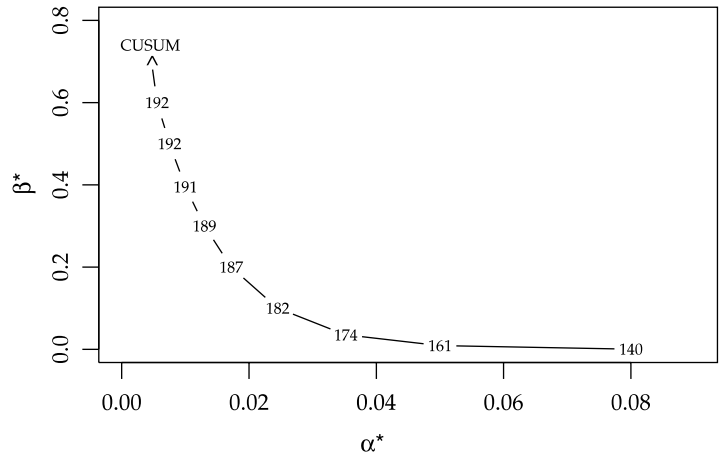

Fig. 3. Change in out-of-control ARL for pairs (α^*,β^*) given an in-control ARL of 6700 patients: RSPRTs monitoring cardiac surgeons (example 1) (α^* and β^* are defining parameters related to a and b via equations (2) and are not connected to the true error rates $\alpha = 1$ and $\beta = 0$ of the charts)

Table 1. Comparison of ARLs and run length standard deviations under hypotheses H_1 (odds = 2) and H_2 (odds = 2 after 1900 observations) for pairs $(\alpha^*, \beta^*) \equiv (a, b)$ and in-control ARL 6700 patients (example 1 data)

α^*	β^*	a	b	*ARL*		*Standard deviation*	
				H_1	H_2	H_1	H_2
CUSUM		0	4.5	193	185	120	129
0.013	0.3	−1.19	3.99	189	189	126	136
0.0352	0.0352	−3.31	3.31	174	210	121	145
0.05	0.009	−4.66	2.99	161	217	135	159
0.08	0.00045	−7.62	2.53	140	274	134	197

The problem with having α^* set high relative to β^*, i.e. having the lower boundary a more extreme than the upper boundary b is that this makes it possible for substantial 'credit' to be built up in the chart. Thus, unlike the CUSUM which has a holding barrier at 0, an RSPRT chart may accumulate credit up to the amount that is needed to cross the lower boundary.

This credit is a problem if the process is not out of control from the start of monitoring, as it would be if the alternative hypothesis H_1 were true, but, rather, it goes out of control after monitoring has been in place for a period of time. Assume, for example, that in the cardiac surgery example the odds of 30-day mortality double after 1900 patients (the lower quartile of the in-control run length distribution). Call this hypothesis H_2. Table 1 gives the out-of-control ARLs after the change in odds under H_1, when the change is immediate, and under H_2, for various pairs $(\alpha^*, \beta^* \equiv (a, b))$. The corresponding standard deviations are also given. The results were obtained from simulations of 1000 runs. Table 1 shows that the ARL under H_2 is greater for the chart with high α^* and low β^* than for that with low α^* and high β^*, although under H_1 this is the other way round (as demonstrated by Fig. 3). The increase in standard deviation under H_2, and also (less noticeably) H_1, as α^* increases shows that charts with a higher α^*

have more variability in run length than those with a low α^*. In view of this, minimization of the out-of-control ARL is not a sensible optimality criterion when RSPRT charts are used for routine monitoring.

In general, the criteria by which charts are compared in any given situation should be chosen with the properties of the process being monitored in mind. For example, the process may be prone to drift out of control, or to change radically at any stage or may fluctuate. Some criteria may be completely inappropriate for choosing optimal charts under such changes.

4. The Shewhart chart

The Shewhart chart, which was developed by Walter Shewhart in the 1920s, simply charts the actual observations (sometimes standardized) of a process. The process is deemed to be out of control when prespecified probability limits are crossed. Usually 99% limits are set (3σ limits on a standard Shewhart chart for normal data) so that only large changes in the process will be detected and the false alarm rate is reduced. Often two-sided control limits are implemented, but a one-sided limit can also be used.

Since the run length is a discrete waiting time, the run length distribution can be assumed to be geometric with mean equal to 1 over the probability that an outcome falls outside the control limits.

For binary data, it is nonsensical to observe whether single observations in isolation cross probability limits. To use a Shewhart chart, the data must be grouped and assumed, for example, to be binomial or normal. Shewhart charts for binomial data are termed p-charts.

For count data, the data can be charted as they are and assumed to be Poisson or negative binomial distributed, for example.

A risk-adjusted version of the Shewhart p-chart for binomial data has been developed by Cook *et al.* (2003) to track grouped binary outcomes in intensive care. There, they simply allow the probability of failure at each group of observations of size n_t to depend on the case mix and calculate the probability limits for that group under the assumption that the distribution of the number of failures at each time point t may be adequately modelled as $\mathcal{N}\{\Sigma_{i=1}^{n_t} p_{it}, \Sigma_{i=1}^{n_t} p_{it}(1-p_{it})\}$, where p_{it} is the expected probability of failure of the ith patient in a group at time t. Here we make a stronger assumption and say that it could be modelled as $\mathcal{B}(n_t, \bar{p}_t)$ where $\bar{p}_t = (1/n_t)\Sigma_{i=1}^{n_t} p_{it}$, as we feel that this may be more accurate for smaller group sizes. Nevertheless, simulations have shown that this binomial approximation performs similarly to the normal approximation for the case mix and group size that were examined in the examples here.

To apply a risk-adjusted Shewhart chart in the case of count data, we make the assumption that the number of failures in a group at time t of size n_t follows $\mathcal{P}(\lambda_t)$ where $\lambda_t = \Sigma_{i=1}^{n_t} \lambda_{it}$, the sum of the individual rates of failure. This result is exact, as long as the data are Poisson distributed. For the binomial case, it is thought that, as long as the distribution of patient type probabilities is tight and non-skew, the approximation is reasonable, but that where the distribution is flat or highly skewed the ARL should perhaps be checked by simulation.

4.1. Example

For the example 1 data set, suppose that we wish to conduct a one-sided test for a doubling in the odds of 30-day mortality, i.e. to test H_0: $p_{r0} = p_r$ *versus* H_1: $p_{r1} = Rp_r/\{1+(R-1)p_r\}, r = 1, 2, \ldots, m$. Now, the failure rate averaged over the Parsonnet score (the factor determining case type), $\bar{p}$, can be calculated from the data set and is taken to be 0.066. For a p-chart with groups of size 79 we wish to test whether the number of failures within 30 days under one surgeon follows $\mathcal{B}(79, \Sigma_{j=1}^{79} p_j/79)$ at the $\alpha\%$ level. The size of group was chosen to be 79 to correspond

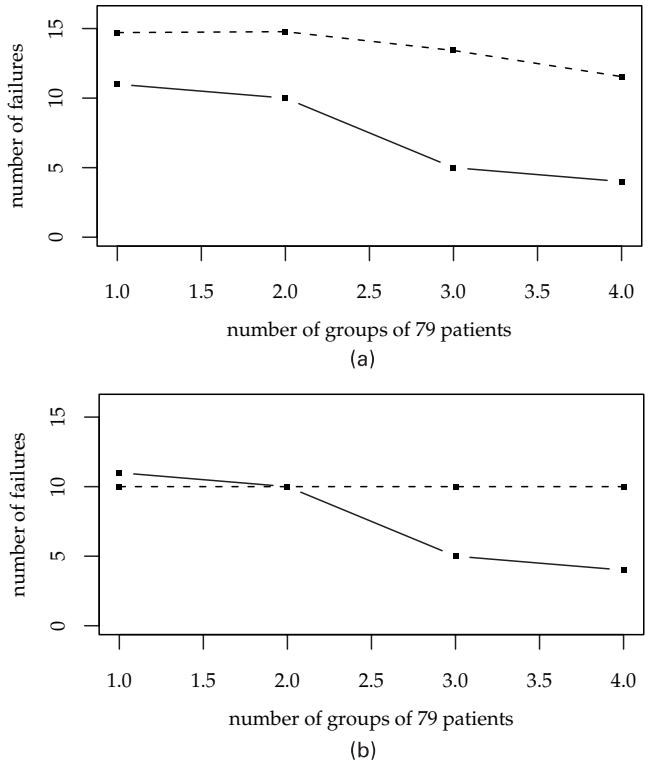

Fig. 4. 6-monthly Shewhart charts testing for a doubling in the odds of 30-day mortality following cardiac surgery (in-control ARL, 84 6-monthly periods): (a) adjusted; (b) unadjusted

roughly to 6 months of surgery for an average surgeon. For the single surgeon considered here, 79 patients correspond more closely to 9 months of surgery.

To achieve an in-control ARL of 6700 (84 sets of 79 patients), α needs to be set at $1/84 = 0.012$. The out-of-control ARL for this chart is 294 (four sets of 79 patients).

Fig. 4 shows a risk-adjusted Shewhart chart for the example 1 data constructed as described. The limit is not crossed, so the odds of 30-day mortality are deemed not to have changed. For comparison, the unadjusted chart is given as well. This has the same in- and out-of-control ARLs as the risk-adjusted chart but has fixed limits calculated under the assumption that the number of failures every 79 patients follows $\mathcal{B}(79, \bar{p} = 0.066)$, where $\bar{p}$ is the failure rate averaged over patient type from the training data. The chart signals at the first group of observations.

5. The sets method and time-between-events Shewhart chart

The sets method for monitoring adverse outcomes was introduced by Chen (1978) as a surveillance system for congenital malformations. Gallus *et al.* (1986) later refined the method. The method was adapted to allow for risk adjustment by Grigg and Farewell (2004).

The risk-adjusted method is based on the 'set' X of *weighted* observations after failure up to and including the next failure, where the weights are dependent on risk. If $X \leqslant T$ on n successive occasions, then an alarm is signalled and the process is deemed to have moved from an initial state H_0 to an out-of-control state H_1. When $n = 1$, the method is equivalent to a one-sided Shewhart-type chart monitoring time between events. The weights that were suggested by Grigg and Farewell (2004) are $p_r/\bar{p}$, where $\bar{p}$ is the probability of failure averaged over the type of risk

and p_r is defined as previously. Using these weights, the weight for an 'average' patient is 1, the weight that is adopted in the unadjusted method.

In the original method, $X \sim \text{geometric}(\pi_0)$ under the null hypothesis, and similarly $X \sim \text{geometric}(\pi_1)$ under the alternative hypothesis. The distribution of X for the risk-adjusted method, however, is intractable and probability calculations involving X must be carried out by using an empirical distribution based on simulations.

Define the event $X \leqslant T$ to be an A-event and a B-event its complement. The probability of an alarm (under either hypothesis) is therefore given by

$$P_i(\text{alarm}) = P_i^n(A) \qquad i = 0, 1. \tag{7}$$

To set the in-control ARL S_0, we require that the probability of an alarm under H_0 satisfies

$$P_i(\text{alarm}) = \lim_{\alpha \to \infty} \left(\frac{\alpha}{\alpha D_i - n + 1} \right) = \frac{1}{D_i} \qquad i = 0, 1 \tag{8}$$

for $i = 0$. For $i = 1$ equation (8) gives the relationship between the probability of a true alarm, $P_1(\text{alarm})$, and the out-of-control ARL S_1. The terms α and $\alpha D_i - n + 1$ correspond to the numbers of actual and possible alarms respectively, and $D_i = \pi_i S_i$ is the number of failures that are expected over S_i patients.

In terms of A-events, $P_i(\text{alarm})$ must also satisfy the relationship

$$P_i(\text{alarm}) = \frac{P_i^n(A)\{1 - P_i(A)\}}{1 - P_i^n(A)} \qquad i = 0, 1. \tag{9}$$

This equation recognizes the fact that alarms are considered as disjoint, i.e. an A-event following n consecutive A-events results in an alarm only if the $n - 1$ previous A-events were not part of a previous alarm.

Equating expressions (8) and (9) gives

$$D_i = \frac{1 - P_i^n(A)}{P_i^n(A)\{1 - P_i(A)\}} \qquad i = 0, 1, \tag{10}$$

which can be rearranged to give

$$P_i(A) = \{1 + D_i - D_i\, P_i(A)\}^{-1/n} \qquad i = 0, 1. \tag{11}$$

Equations (10) and (11) are used to find values for n and T such that D_1 is minimal or, equivalently, so that the out-of-control ARL D_1/π_1 is minimal. According to Gallus *et al.* (1986) simulation results suggest that D_1 does have a unique minimum with respect to n and T. The values for n and T are found by an iterative procedure that can be terminated as soon as the value of D_1 is found to be higher than at the previous iteration.

The iterative procedure is as follows, starting with $n = 2$.

(a) Calculate the value of $P_0(A)$ by applying Newton–Raphson iteration to equation (11), using as an initial value the solution to equation (11) obtained for $n - 1$.
(b) Interpolate the value of T from the simulated empirical distribution of X under H_0.
(c) Interpolate the value of $P_1(A)$ from the simulated empirical distribution of X under H_1.
(d) Calculate D_1 from equation (10).
(e) If $n + 1 > D_0$, stop.
(f) Increase n by 1.

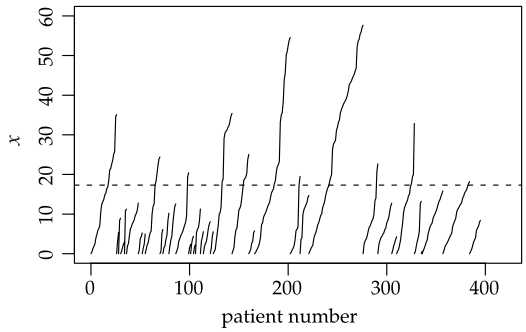

Fig. 5. Grass plot retrospectively monitoring a cardiac surgeon (example 1; $n = 10$; $T = 17.31$)

5.1. Example

Suppose that we want to test (for the example 1 data) H_0: $p_{r0} = p_r$ *versus* H_1: $p_{r1} = Rp_r/\{1 + (R-1)p_r\}$, $r = 1, 2, \ldots, m$. Assuming that the failure rate averaged over the Parsonnet score is 0.066, for a patient of type r we would add the weight $p_r/0.066$ to X.

By simulating the in- and out-of-control distributions of X under hypotheses H_0 and H_1 respectively, and following the algorithm above with an in-control ARL fixed at 6700 patients, the optimal values for n and T were found to be 10 and 17.31. The corresponding out-of-control ARL was 324.6. Since the optimal value for n is not equal to 1, the Shewhart time between events chart is not optimal for this particular data set. Fig. 5 illustrates a sets chart (grass plot), as proposed by Grigg and Farewell (2004), with values of n and T that are appropriate for the example 1 data. The chart simply plots the observation number against the cumulative size of the current set or blade. The chart is reset to 0 (a new set is begun) after every failure.

6. Comparison of charts for binary data

Chen (1987) suggested that the original sets method performs better than the CUSUM when the rate of adverse outcomes is low. Gallus *et al.* (1986) and Barbujani and Calzolari (1984) questioned the results of Chen, but Gallus *et al.* (1986) argued that their modified sets method can be more efficient than the CUSUM, but under different circumstances. The examples that they gave demonstrate that the refined sets method performs better when the change in rates that it is designed to detect is large, and not necessarily when the initial rate is low.

Here we have compared the sets method with the CUSUM and also the Shewhart p-chart, using the case mix from example data set 1 as a basis for the comparison. The focus of the comparison is on two factors: the size of the change in parameter being tested for and the case mix probabilities. So, the relative efficiencies of the charts under testing for a doubling in odds of 30-day mortality compared with testing for a fivefold increase were calculated. This was done for each of three sets of case mix probabilities: the original probabilities ($\bar{p} = 0.066$); the original probabilities multiplied by $\lambda = 0.5$ ($\bar{p} = 0.033$); the original probabilities multiplied by $\lambda = 1.5$ ($\bar{p} = 0.099$).

Fig. 6 shows the ratio of the out-of-control ARL for the sets method to that for the CUSUM measured for various values of in-control ARL (on a log-scale).

For charts testing for a larger increase in parameter (Fig. 6(b)) it appears that the sets method is more efficient (for the smallest two of the three sets of case mix probabilities) than the CUSUM for a slightly larger range of in-control ARL values than when the increase in the parameter is smaller (Fig. 6(a)). By efficient, it is meant that the out-of-control ARL is smaller for a

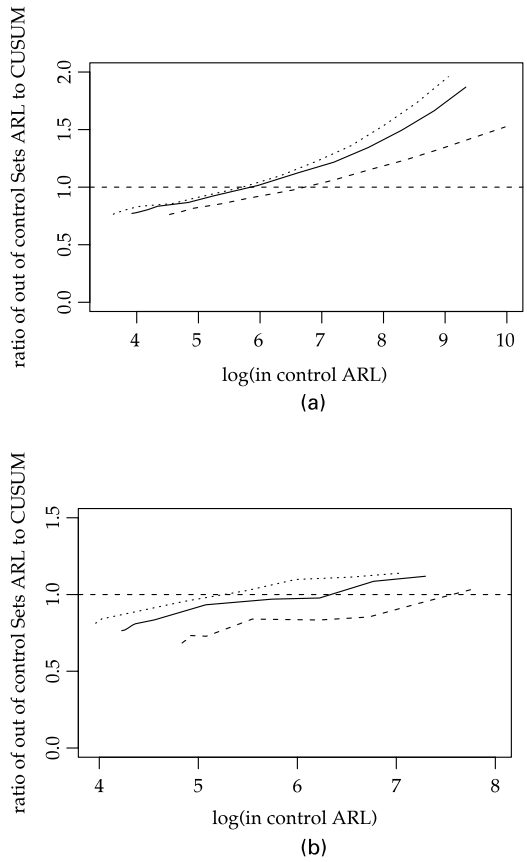

Fig. 6. Ratio of out-of-control ARL for the sets method to that of the CUSUM for fixed log(in-control ARL) (———, $\lambda = 1$; -------, $\lambda = 0.5$; ·····, $\lambda = 1.5$): (a) charts testing for a doubling in odds; (b) charts testing for a fivefold increase in odds

fixed in-control ARL. The size of case mix probabilities seems to have a noticeable effect. The larger the case mix probabilities, the less efficient the sets method is compared with the CUSUM. Although the effect can be seen, it does not appear to be large. However, this may be because the change in average case mix probabilities across the three sets (0.033,0.066,0.099) is small.

Tables 2 and 3 give the full results, including the Shewhart chart ARLs. Table 2 compares charts testing for a doubling of the odds of 30-day mortality following cardiac surgery and Table 3 compares charts testing for a fivefold increase.

From both Tables 2 and 3 we see that the CUSUM chart is uniformly better than the Shewhart p-chart. For charts testing for a smaller change in odds ratio (Table 2) this is not surprising. However, for charts testing for a larger change in odds ratio (Table 3), the Shewhart chart might be expected to be more efficient than the CUSUM. The result is thought to be because observations must be grouped for the Shewhart chart. In the case of the larger change in odds (Table 3), the sets method is uniformly better than the Shewhart p-chart. In the case of the smaller change in odds (Table 2), for larger in-control run lengths (log(in-control ARL) > 8) and larger case mix probabilities ($\lambda = 1$ and $\lambda = 1.5$), the Shewhart chart is more efficient than the sets method, especially for $\lambda = 1.5$.

Table 2. Out-of-control ARLs of the sets method, CUSUM and Shewhart (group size 79) charts for a fixed in-control ARL: charts testing for a doubling in odds of 30-day mortality following cardiac surgery

λ	*log(in-control ARL)*	*ARLs for the following methods:*			λ	*log(in-control ARL)*	*ARLs for the following methods:*		
		Sets	*CUSUM*	*Shewhart*			*Sets*	*CUSUM*	*Shewhart*
0.5	4.62	31.4	40.6	158	1.5	3.74	13.8	17.5	79
	4.74	34.7	43.9	158		3.88	15.7	19.3	79
	4.89	39.3	48.3	158		4.06	18.1	21.7	79
	5.38	56.6	66.1	158		4.58	25.8	30.1	158
	5.80	76.2	84.4	158		4.96	34.0	37.5	158
	6.54	121	124	237		5.69	54.4	54.6	158
	7.19	175	165	237		6.33	79.9	72.7	158
	7.79	240	208	395		6.92	112	91.4	158
	8.36	313	253	395		7.47	151	111	158
	8.91	395	297	632		8.01	200	130	158
	9.45	488	342	632		8.54	258	149	237
	9.97	591	388	1264		9.05	331	169	237
1	4.04	18.3	23.3	79					
	4.18	20.6	25.6	79					
	4.34	23.6	28.3	79					
	4.84	33.8	39.0	158					
	5.24	45.2	49.0	158					
	5.97	72.6	71.5	158					
	6.61	107	95.3	158					
	7.20	146	120	158					
	7.76	195	145	237					
	8.29	254	170	237					
	8.82	326	196	316					
	9.34	415	222	316					

7. Testing for improvements

When a process is to be monitored long term to detect a deterioration in the process, it is also important to take note of improvements in the process. If improvements in the process are ignored, a chart may be less sensitive to subsequent deteriorations in the process.

For the joint monitoring of improvement and deterioration, Page (1954) suggested the use of a two-sided CUSUM, i.e. the combined use of two one-sided tabular CUSUMs: one to detect improvement; one to detect deterioration. However, calculation of the ARL was not demonstrated.

More recently, Khan (1984) investigated the relationship between the run length of two one-sided CUSUMs, upper and lower, and a single two-sided CUSUM. The approximate formula derived was

$$\frac{1}{\mathrm{ARL}^{\mathrm{c}}} = \frac{1}{\mathrm{ARL}^{+}} + \frac{1}{\mathrm{ARL}^{-}} \tag{12}$$

under certain regularity conditions, where $\mathrm{ARL}^{\mathrm{c}}$ is the ARL for the two-sided CUSUM, ARL^{+} is the ARL for the upper one-sided CUSUM and ARL^{-} is the ARL for the lower one-sided CUSUM. Intuitively, the formula represents an assumption that the two sides of the chart are independent (the regularity conditions are, essentially, that the two halves cannot interact). The

Table 3. Out-of-control ARLs of the sets method, CUSUM and Shewhart (group size 40) charts for a fixed in-control ARL: charts testing for a fivefold increase in odds of 30-day mortality following cardiac surgery

λ	*log(in-control ARL)*	*ARLs for the following methods:*			λ	*log(in-control ARL)*	*ARLs for the following methods:*		
		Sets	*CUSUM*	*Shewhart*			*Sets*	*CUSUM*	*Shewhart*
0.5	4.84	11.3	16.5	80	1.5	3.96	6.47	7.95	40
	4.87	11.6	16.6	80		3.99	6.62	8.05	40
	4.89	11.8	16.8	80		4.03	6.93	8.13	40
	4.92	12.5	17.0	80		4.07	7.01	8.28	40
	5.08	13.2	18.1	80		4.29	7.66	9.08	40
	5.55	18.3	21.8	80		4.86	10.7	11.5	40
	6.22	23.2	27.8	80		5.48	14.6	14.2	40
	6.72	27.5	32.2	80		5.98	18.2	16.6	40
	7.24	35.2	37.2	80		6.53	21.2	19.2	40
	7.79	44.1	42.5	80		7.05	24.8	21.8	80
1	4.23	7.62	9.97	40					
	4.26	7.75	10.1	40					
	4.29	7.94	10.2	40					
	4.33	8.31	10.4	40					
	4.36	8.54	10.5	40					
	4.56	9.47	11.4	40					
	5.08	13.1	14.0	40					
	5.74	17.2	17.7	40					
	6.22	20.0	20.5	80					
	6.77	25.8	23.8	80					
	7.29	30.2	27.0	80					

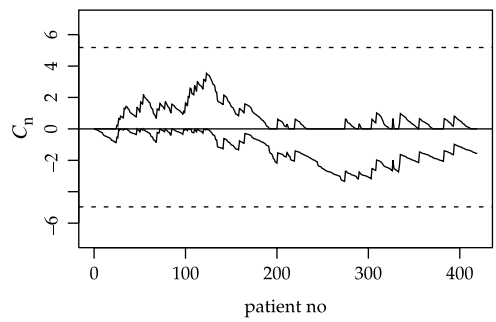

Fig. 7. Two-sided risk-adjusted CUSUM with $h_u = 5.18$, $h_l = -4.96$ and in-control ARL 6700 patients (example 1 data)

details are not shown here, but a simulation study demonstrated that this formula also applies to RSPRTs and that it works equally well for CUSUMs and RSPRTs with risk adjustment.

Fig. 7 demonstrates a two-sided risk-adjusted CUSUM for the example 1 data. The upper and lower boundaries have been chosen to be $h_u = 5.18$ and $h_l = -4.96$ respectively, so that the in-control ARLs for each half of the chart are both equal to 13400. This means that the overall ARL is approximately $13400/2 = 6700$. The in-control ARLs on each side have been made the same here to balance out the false alarm rate. However, we might decide to allow for more false positive than negative alarms, or vice versa, in which case an asymmetric chart should be employed.

8. Estimation

The primary purpose of a chart is not to estimate a process parameter. Even so, it is often natural to want to provide some estimate of a parameter after a warning signal.

Consider an unadjusted CUSUM chart for the data of example 2, where the rate of deaths per year, λ, is the parameter of interest. The difficulty, for a frequentist analysis at least, is that the maximum likelihood estimate (MLE) $\hat{\lambda}$ is biased. Although the likelihood is not affected by the stopping rule, the sampling distribution of $\hat{\lambda}$ is.

An approach to the problem, suggested by Grigg *et al.* (2003), is to obtain an MLE and then to implement Whitehead's (1997) method for adjusting the bias. This approach involves finding the bias function $b(\lambda)$ for a particular chart and solving

$$\tilde{\lambda} = \hat{\lambda} - b(\tilde{\lambda}) \tag{13}$$

where $\hat{\lambda} = Y_n/n$ is the MLE of λ at the point of stoppage, n. If the bias function is difficult to attain explicitly, a simulated approximation can work just as well.

For risk-adjusted charts, it is easier to deal with $\bar{\lambda}$ when constructing bias curves, where $\bar{\lambda}$ is the failure rate (per year) averaged over the type of patient, than to have multiple bias curves, one for each patient type's rate.

An estimate taken from a chart can be based on all the observations, since the start of monitoring. However, it is common, also, to base estimates on only the data that are observed after the estimated 'changepoint'. This is the point at which the process is deemed to have moved from a null state to an out-of-control state.

For RSPRTs, the time that the chart was last at a, the lower boundary, is the estimate of the changepoint (extended from the result stated for CUSUMs by Hawkins and Olwell (1997)). For the sets method it is estimated as being the observation before the start of the last n consecutive sets.

8.1. Example

Consider, for the example 2 data, a two-sided CUSUM testing the null hypothesis H_0: $\lambda_0 = 35$ *versus* the alternative H_u: $\lambda_u = 1.2\lambda_0$, H_l: $\lambda_l = 0.8\lambda_0$. Now, $\Pr\{\mathcal{P}(42) > 150\} \approx 0$. Therefore, constraining Y_t, the number of deaths per year, to be 150 or fewer should not result in much loss of information. From Section 3, the log-likelihood ratio weights are

$$\left.\begin{array}{l} W_t(u) = Y_t \log(1.2) - 7, \\ W_t(l) = Y_t \log(0.8) + 7 \end{array}\right\} \qquad Y_t \in \{0, 1, 2, \ldots, 150\} \tag{14}$$

where u and l refer to the upper and lower charts respectively.

Fig. 8 shows a two-sided CUSUM consisting of these weights monitoring the observed death-rates per year over all types of patient under Harold Shipman, 1987–1998. Boundaries have been arbitrarily placed at $h = 1, 2, 3, 4, 5$ on both sides.

The chart would signal at the end of 1995 if h were chosen to be anywhere in the range [3,8] because of the extreme increase in rate after 1994. For boundaries at (−3,3), the in-control ARL is 52 years and the out-of-control ARL 5 years; at (−5,5), they are 403 and 7.5 years respectively.

Applying Whitehead's method of bias adjustment to the MLE of λ at the end of 1995, calculated from all the data, results in an adjusted value of $\lambda = 41$ from $\hat{\lambda} = 42.33$. An approximate 95% confidence interval for λ is [37,45].

The MLE since the chart was last at 0 (at the end of 1992) is $\hat{\lambda} = 53.33$. Applying Whitehead's method gives an adjusted value of $\lambda = 52$. An approximate 95% confidence interval for

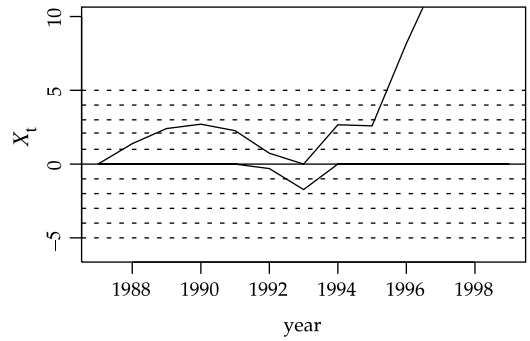

Fig. 8. CUSUM monitoring death-rates per year under Harold Shipman, 1987–1998

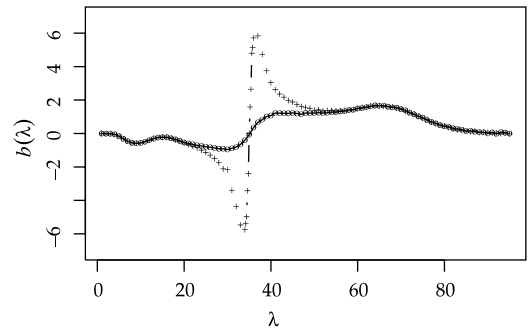

Fig. 9. Simulated bias curves for a two-sided CUSUM with $\lambda_0 = 35$, $\lambda_u = 1.2\lambda_0$, $\lambda_l = 0.8\lambda_0$ and $h_u = h_l = 5$ (example 2 data): ∘, all data; +, data since last at 0

this adjusted estimate is [43,60]. Note that the obtained confidence interval barely overlaps the interval that is obtained for the estimate using all the data (since 1987).

Fig. 9 illustrates the simulated bias curve for the MLE calculated by using all the data for a two-sided CUSUM chart with $\lambda_0 = 35$, $\lambda_u = 1.2\lambda_0$, $\lambda_l = 0.8\lambda_0$ and $h_u = h_l = 5$. The bias curve relating to estimates calculated by using only the data since the chart was last at 0 is also given.

The fact that the choice of estimator for λ results in such different values is evidently a problem. If a change in the process occurred in 1992, we would like to estimate the parameter by using data from that point onwards only. However, if a change had not occurred, or occurred earlier than 1992, not using the earlier data to form the estimate might result in an estimate with a large bias. Even if a bias adjustment were made, considerable bias could still remain.

There is an assumption here, also, that a change in process, if it occurs, will be immediate and sustained. Changes, in practice, however, might be gradual, or intermittent. In this case, other estimators than the two described might be more appropriate.

9. Conclusion

A variety of risk-adjusted charts have been presented here. Comparisons between the charts are based on the empirical case mix distribution from a single data set. However, it is thought that the results could, with due caution, be generalized to a broader spectrum of data. More work in this area, on other contrasting data sets, is certainly required for a greater understanding of how the methods presented compare.

For the RSPRT class of charts (which includes the CUSUM as a special case) it is shown that the optimal chart, under the 'minimum out-of-control ARL for a fixed in-control ARL' criterion is an RSPRT with low α^* and high β^*. It is argued, however, that this criterion is not sensible for optimizing over this class, because the optimal chart chosen is the chart that can build up the most credit and therefore is the least sensitive to changes in the process that occur at any time other than early on in monitoring.

A comparison is also made between the sets method, the CUSUM and the Shewhart p-chart. For the sets method and CUSUM, the aim was to broaden and clarify comparisons of the two charts that have been made previously. From the results gathered, it is recommended that, when wishing to detect small changes in a low event rate process, the sets method should be used only if the changes need to be detected extremely quickly regardless of a higher rate of false alarms. Otherwise, the CUSUM is perhaps the better tool. The size of the underlying case mix probabilities has a clear but relatively small effect on the comparative efficiency of the charts in the example that was studied. Because of the constraints of the data set, though, it is difficult to say whether this effect might be more significant for larger changes in the case mix probabilities.

The Shewhart chart is included in the comparison, because it is a standard and simple chart. The chart is found (for these data, at least) to be less efficient than the CUSUM. This is thought to be because, to monitor binary data, it must necessarily work on groupings of the data. For charts testing for larger changes in parameters, and for charts with smaller in-control ARLs otherwise, it is also found to be less efficient than the sets method.

The importance of implementing charts that can detect improvement as well as deterioration in a process is highlighted. Taking note of improvement may prompt a re-evaluation of the standard and, moreover, identify centres or individuals who perform well. If those centres or individuals have a transferable method of working, a positive feed-back system could be induced.

With regard to estimating the process parameter from a chart, rather than using the chart directly, the parameter could be estimated by an on-going smoothing process, such as a straightforward exponentially weighted moving average, a Bayesian exponentially weighted moving average or possibly by the use of full Bayesian updating. Indeed, if estimation is of central importance, rather than quality control, using such techniques might be an alternative strategy to the use of control charts.

Concerning the practical issue of which charts retrospectively would have been best implemented to monitor the surgical data (example 1), the results of the comparison of Section 6 suggest that either the sets method or CUSUM are most efficient for the particular case mix that was observed: the rate of deaths in the charted data (1994–1998) is 0.086, corresponding to $\lambda = 1.3$. If we had wanted to detect a change in rate quickly or to detect only large changes in rate, the sets method would probably have been the more suitable method. However, the CUSUM would perhaps have been easier to implement.

For the Shipman data, any chart monitoring the rate of deaths among elderly females would have been useful for an early detection of the problem (Fig. 2, for example, illustrates that there were over 20 excess deaths (going by the averages for England and Wales) of females over 75 years per year going back to the end of 1988—the results of the inquiry suggest that about half of these were probably caused intentionally by Shipman). However, the prospective identification by control chart of problems in such subgroups, and indeed problems that are specific to one general practitioner among many, would prove difficult for two reasons. Firstly, using charts to monitor several subgroups simultaneously would mean a loss of power for each individual chart concerned, owing to the multiplicity. Secondly, the run length properties of a large group of combined charts are as yet unknown.

References

Baker, R. (2001) *Harold Shipman's Clinical Practice 1974–1998: a Review commissioned by the Chief Medical Officer*. London: Stationery Office.

Barbujani, G. and Calzolari, E. (1984) Comparison of two statistical techniques for the surveillance of birth defects through a monte carlo simulation. *Statist. Med.*, **3**, 239–247.

Chen, R. (1978) A surveillance system for congenital malformations. *J. Am. Statist. Ass.*, **73**, 323–327.

Chen, R. (1987) The relative efficiency of the sets and the cusum techniques in monitoring the occurrence of a rare event. *Statist. Med.*, **6**, 517–525.

Cook, D. A., Steiner, S. H., Farewell, V. T. and Morton, A. P. (2003) Monitoring the evolutionary process of quality: risk adjusted charting to track outcomes in intensive care. *Crit. Care Med.*, **31**, 1676–1682.

DeLeval, M. R., François, K., Bull, C., Brawn, W. B. and Spiegelhalter, D. J. (1994) Analysis of a cluster of surgical failures. *J. Thor. Cardvasc. Surg.*, **104**, 914–924.

Gallus, G., Mandelli, C., Marchi, M. and Radaelli, G. (1986) On surveillance methods for congenital malformations. *Statist. Med.*, **5**, 565–571.

Grigg, O. A. and Farewell, V. T. (2004) A risk-adjusted Sets method for monitoring adverse medical outcomes. *Statist. Med.*, to be published.

Grigg, O. A., Farewell, V. T. and Spiegelhalter, D. J. (2003) Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Statist. Meth. Med. Res.*, **12**, 147–170.

Hawkins, D. M. and Olwell, D. H. (1997) *Cumulative Sum Charts and Charting for Quality Improvement*. New York: Springer.

Khan, R. A. (1984) On cumulative sum procedures and the SPRT with applications. *J. R. Statist. Soc.* B, **46**, 79–85.

Lovegrove, J., Sherlaw-Johnson, C., Valencia, O., Treasure, T. and Gallivan, S. (1999) Monitoring the performance of cardiac surgeons. *J. Oper. Res. Soc.*, **50**, 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*, **350**, 1128–1130.

Page, E. S. (1954) Continuous inspection schemes. *Biometrika*, **41**, 100–115.

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*, **79**, suppl. 1, 1–12.

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

Shipman Inquiry (2002) *Shipman Inquiry: the First Report*. London: Stationery Office. (Available from `http://www.the-shipman-inquiry.org.uk/firstreport.asp`.)

Spiegelhalter, D. J., Grigg, O. A., Kinsman, R. and Treasure, T. (2003) Sequential probability ratio tests (sprts) for monitoring risk-adjusted outcomes. *Int. J. Qual. Hlth Care*, **15**, 1–7.

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

Wald, A. (1945) Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, **16**, 117–186.

Whitehead, J. (1997) *The Design and Analysis of Sequential Clinical Trials*, 3rd edn. Chichester: Horwood.