OPEN ACCESS

Comparison of control charts for monitoring clinical performance using binary data

Jenny Neuburger,[1,2] Kate Walker,[2,3] Chris Sherlaw-Johnson,[1] Jan van der Meulen,[2,3] David A Cromwell[2,3]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2016-005526).

[1]The Nuffield Trust, London, UK
[2]Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, Greater London, UK
[3]Clinical Effectiveness Unit, The Royal College of Surgeons of England, London, UK

Correspondence to
Dr Kate Walker, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK;
kate.walker@lshtm.ac.uk

JN and KW contributed equally.

Received 24 March 2016
Revised 24 May 2017
Accepted 30 May 2017
Published Online First
25 September 2017

To cite: Neuburger J, Walker K, Sherlaw-Johnson C, *et al*. *BMJ Qual Saf* 2017;**26**:919–928.

ABSTRACT

Background Time series charts are increasingly used by clinical teams to monitor their performance, but statistical control charts are not widely used, partly due to uncertainty about which chart to use. Although there is a large literature on methods, there are few systematic comparisons of charts for detecting changes in rates of binary clinical performance data.

Methods We compared four control charts for binary data: the Shewhart p-chart; the exponentially weighted moving average (EWMA) chart; the cumulative sum (CUSUM) chart; and the g-chart. Charts were set up to have the same long-term false signal rate. Chart performance was then judged according to the expected number of patients treated until a change in rate was detected.

Results For large absolute increases in rates (>10%), the Shewhart p-chart and EWMA both had good performance, although not quite as good as the CUSUM. For small absolute increases (<10%), the CUSUM detected changes more rapidly. The g-chart is designed to efficiently detect decreases in low event rates, but it again had less good performance than the CUSUM.

Implications The Shewhart p-chart is the simplest chart to implement and interpret, and performs well for detecting large changes, which may be useful for monitoring processes of care. The g-chart is a useful complement for determining the success of initiatives to reduce low-event rates (eg, adverse events). The CUSUM may be particularly useful for faster detection of problems with patient safety leading to increases in adverse event rates.

INTRODUCTION

There is growing interest in using time series charts for monitoring clinical performance in order to assess safety and improve quality of care.[1–4] However, this way of displaying data can lend itself to overinterpretation of chance fluctuations, potentially resulting in inappropriate decisions and improvement fatigue.[5 6] Control charts with statistical control limits are useful for distinguishing *systematic* changes from *chance* variation in processes of care and outcomes.[7]

Although well used by some organisations, control charts are still not widely used to monitor quality and safety in routine healthcare delivery.[8–12] For example, a review of 1488 charts used by 30 English hospital boards found that just 6% of charts included limits to depict the role of chance.[8] Uncertainty about which chart to use has been identified as a barrier to the use of control charts in clinical settings.[8 9] The myriad choices to be made when setting up more complex charts may also limit their accessibility.[3]

Clinical performance is often measured using indicators derived from binary event data such as care meeting a specified standard, or mortality. Shewhart charts for binary data include the p-chart that tracks the proportion with an event for consecutive periods[13] and the g-chart that displays the number of cases between events and is specifically designed to detect reductions in event rates.[14] More complex charts that accumulate information over time include the exponentially weighted moving average (EWMA) chart and the cumulative sum (CUSUM) chart. These can detect small increases in event rates more quickly than the p-chart.[15–17]

Although there are several excellent reviews of different charts used in clinical settings,[4 16] researchers have not systematically reviewed how quickly different charts detect change over a range of effect sizes that are realistic for binary clinical data.[7] This contrasts with the industrial statistical process control (SPC) literature, in which such comparisons are standard.[18]

In this paper, we compare four control charts for binary clinical data: the Shewhart p-chart;[13] the g-chart;[14] the EWMA chart;[19] and the CUSUM chart.[20–22] We describe how to set up these charts to detect increases and decreases in rates. We then compare the charts based on the expected number of patients until a change in the event rate is detected, referred to as the average run

length (ARL). We also describe the impact of altering chart settings on ARLs.

EXAMPLE OF CONTROL CHARTS FOR MONITORING MORTALITY

The control charts are shown in figure 1. They use an example of local monitoring of 90-day mortality after major resection for bowel cancer in one hospital.[23] The average number of elective patients undergoing major resection for bowel cancer was 30 per month in this hospital and baseline mortality was 3%.

The formulae for the charts and their control limits are given in table 1.

In the Shewhart p-chart, data are aggregated, in this case by month, and the chart statistic is the proportion of patients undergoing a major resection who died within 90 days.

The g-chart monitors the number of patients who survived their procedure between each death, and a point is plotted for each patient who died. The trace moves in the opposite direction to the other charts: an increase in numbers of procedures between deaths corresponds to a reduction in mortality. The closely related t-chart can be used to monitor time between events, such as days.

The EWMA chart statistic is a weighted moving average of current and past individual outcomes and is updated with each procedure. The weight is exponential, meaning that the contribution of past observations decreases going back in time. Like the p-chart

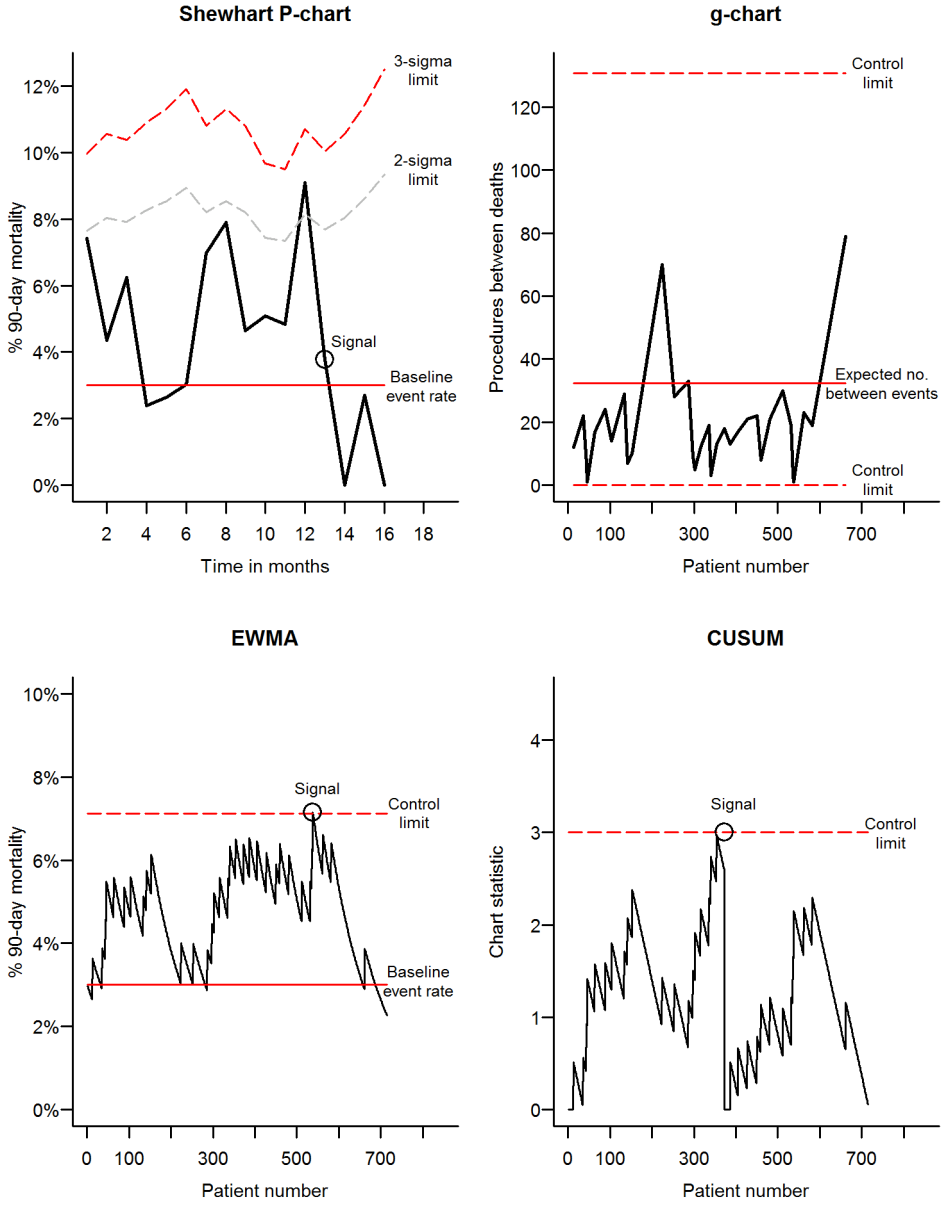

Figure 1 Comparison of four control charts for local monitoring of 90-day mortality following major resesection for bowel cancer in one hospital. CUSUM, cumulative sum; EWMA, exponentially weighted moving average.

Table 1 Chart formulae and settings used in chart comparisons

	Shewhart p-chart	g-chart	EWMA	CUSUM
Chart statistic	$p_j = \frac{1}{n}\sum_{i=1}^{n_j} x_{ij}$ where p_j is the proportion in the j^{th} reporting period, n_j is the volume and x_{ij} is the i^{th} observation in the j^{th} period	$S_k = i_k - i_k - 1$ where s_k is the number of observations between events, i_k is the sequence number of the kth event	$S_i = \lambda x_i + (1-\lambda)S_{i-1}$ where s_i is the i^{th} value of the trace, x_i is the ith observation and λ is the weighting parameter	$S_i = max(0, S_{i-1} + w_i)$ where $w_i = \log OR - \log(1 + p(OR - 1))$ *if* $x_i = 1$ $w_i = -\log(1 + p(OR - 1))$ *if* $x_i = 0$ where s_i is the ith value of the trace, x_i is the ith observation, p is the baseline event rate and *OR* is the OR corresponding to the minimum shift size the chart should detect
Target line	Baseline event rate, p	Baseline mean number between events, $\bar{s}$ or (1−p)/p	Baseline event rate, p	No target line. Trace moves between zero and the control limit
Control limits	$p \pm L \times \sqrt{\frac{p(1-p)}{n}}$ where L is the number of SEs (sigma) from the target line (see main text for runs rules)	$\frac{1-p}{p} \pm L\sqrt{\frac{1-p}{p^2}}$	$p \pm L\sqrt{\frac{p(1-p)\lambda\left[1-(1-\lambda)^{2i}\right]}{2-\lambda}}$	Absolute value for limit
Chart settings	p: <0.02 / 0.02 to <0.1 / 0.1 to 0.5; n per period: 200 / 100 / 50	Only limits are altered	λ=0.01	*OR*=2 for an increase and 0.5 for a decrease
In-control ARLs*	For L=3 with runs rules, n as above: p 0.01 / 0.10 / 0.50; In-control ARL 14 400 / 9 800 / 8 750	To detect a decrease: p 0.01 / 0.10 / 0.50; L 4.0 / 6.0 / 8.0; In-control ARL 14 800 / 10 500 / 8 200	For λ=0.01: p 0.01 / 0.10 / 0.50; L 4.3 / 3.3 / 2.9; In-control ARL 15 250 / 10 050 / 8 460	For OR=2: p 0.01 / 0.10 / 0.50; Limit 3.5 / 3.5 / 6.0; In-control ARL 14 290 / 9 300 / 8 900

*In-control ARLs are given for specified values of control limits. Control limits were more finely adjusted to achieve in-control ARLs for comparisons in table 2.
ARL, average run length; CUSUM, cumulative sum; EWMA, exponentially weighted moving average.

statistic, it provides an estimate of current 90-day mortality.

The CUSUM chart statistic is a log-likelihood ratio summarising the evidence that mortality has shifted away from the baseline rate of 3% to a specified alternative event rate, in this case, 5%. It lacks the direct interpretability of the other chart statistics, but the important feature is that a higher value corresponds to stronger evidence.

Interpretation of control limits

In each chart, control limits define the region within which the chart statistic is expected to lie if current mortality is consistent with the baseline (expected) rate of 3%, subject only to chance variation.

The Shewhart p-chart is set up with 3-sigma limits, where sigma is the SD of the grouped monthly data (see online supplementary appendix for the formula). It also uses supplementary runs rules related to the 2-sigma limits. Any of the following would typically be considered evidence of a change in the event rate: one point outside the 3-sigma limits; two out of three consecutive points outside the 2-sigma limits; and eight consecutive points always above or always below the baseline event rate.[13]

The Shewhart p-chart statistic does not cross either of the 3-sigma control limits during the 16 months of monitoring, but a run of 8 points exceeds the baseline event rate at month 13 signalling an increase in mortality. The EWMA crosses the upper control limit in month 12 (approximately 500th resection) and the CUSUM triggers in month 9 (approximately 350th resection). The g-chart does not signal a reduction in mortality during this period.

The CUSUM chart statistic takes a minimum value of zero and resets, typically to zero, after the control limit is exceeded. Monitoring then starts again and subsequent evidence about performance is freshly accumulated. The other charts do not require such resetting after a signal; their traces continue to move through the chart space, in or outside the control limits.

Further details about chart settings are given below, and summarised in table 1, and the web supplementary appendix, including the R code used to produce these charts.

METHODS

Using ARL to compare charts

An ideal chart would take a short time to signal a genuine change in performance and a long time to falsely signal a change. The four charts are compared according to the expected number of patients (or procedures) until a change in the event rate is detected. This is commonly known as the out-of-control ARL, a term borrowed from SPC methods developed for quality control of manufactured products.[4 7 15] The term ARL is sometimes used to mean the numbers of groups until a signal for the p-chart,[18] or number of events for the g-chart.[14] In turn, what we call the ARL is also sometimes termed the average number of observations until signal.[18]

When used to monitor a naturally variable process, even when the event rate is unchanged, the chart statistic will eventually exceed the control limits and falsely signal a change. The expected number of observations until a false alarm is signalled is known as the in-control ARL.

Each of the charts can be set up so that they quickly detect changes in performance by reducing the in-control ARL, but this comes at the cost of increasing the rate of false alarms. For this reason, we compare chart types with settings designed to give a similar in-control ARL. We use the term efficient to describe a chart that has a smaller out-of-control ARL for a similar in-control ARL.[16]

For chart comparisons, the charts were set up with an in-control ARL of approximately 10 000 individual observations until the relevant control limit was exceeded. This corresponds to a low false alarm rate, with one false signal expected every 10 000 patients or procedures. This is suitable for public monitoring of outcomes such as mortality. For internal monitoring of a process of care measure, charts may be set up with shorter in-control ARLs, since higher false alarm rates may not be such a problem.

Methods for estimating ARLs for each of these types of chart are given in the web supplementary appendix, including the R code.[24 25]

Scope of comparisons

We have compared ARLs of one-sided charts designed to detect either increases or decreases in the event rate. The Shewhart p-chart and the EWMA chart can be used to detect reductions in event rates as long as the lower control limit is not at zero. In contrast, the g-chart is specifically designed to efficiently detect decreases in event rates, and can only detect increases when the lower control limit is not at zero. The charts are complementary in this respect, and although each can theoretically be designed to detect both increases and decreases by narrowing control limits or increasing volumes for the p-chart, this may not be possible in practice or may give in-control ARLs that are very short.

The CUSUM is designed to be one-sided, testing either for a shift towards a specified ‘poor performance’ rate or an ‘improvement’ rate. A single two-sided CUSUM can be set up by using two separate charts in parallel.

The in-control ARLs for two-sided charts can be easily approximated from the ARLs of each one-sided chart.[16 24]

We compare the ARLs of charts across a range of shift sizes that are realistic for binary clinical data. We assume that the historical event rate is known. We express shift sizes as odds ratios (ORs)[26] and absolute changes in rates.

Absolute sizes of shift will be low for rarer clinical outcomes such as the occurrence of a surgical site infection and much higher for binary process-of-care measures such as receiving a diagnostic assessment or treatment within a specified time frame.

For a baseline event rate of 1%, a problem in performance may increase the event rate by only 1%–2%, which is small in absolute terms but corresponds to an OR of 2 to 3. In contrast, for a baseline rate of 50%, a change in practice may change the event rate by as much as 15%–25%, which is large in absolute terms but again corresponds to an OR of approximately 2 to 3.

Setting up the charts

For each of the charts, choices need to be made about the settings of the chart parameters. These choices will influence the behaviour of the chart. The settings used for chart comparisons are described below, and summarised in table 1. The impact of altering these settings is then explored.

The Shewhart p-chart

For chart comparisons, volumes per period are set to be 200 patients for event rates of less than 2%, 100 for event rates of 2% to less than 10%, and 50 for event rates of 10% to 50%, as these give in-control ARLs of approximately 10 000.

Because the aim is primarily to detect sustained changes in performance, rather than transitory variations, we have used the Shewhart p-chart with supplementary runs rules outlined above.

The g-chart

Because the chart is based on the geometrical distribution, which is highly skewed, standard 3-sigma control limits typically translate into short in-control ARLs for detecting reductions in rates (increases in numbers of observations between events). One option is to widen limits to give a desired probability of a false alarm[14] or specified in-control ARL. To achieve an in-control ARL of 10,000 for the purposes of comparison to the other charts, we widened the g-chart control limits to between 4 and 8-sigma limits give the same in-control ARLs as the p-chart (table 1).

The EWMA chart

To construct the EWMA, it is necessary to select a value for the weight known as lambda, λ. The weight, λ, determines how much the EWMA statistic reflects the current observation and how much it reflects the previous observations, with the weight given to an observation j time points earlier than the current observation being $\lambda(1-\lambda)^j$. For example, for a value for λ of 0.01, the current observation is given weight 1% and the weights assigned to observations 10, 50 and 100 time points earlier are 0.9%, 0.6% and 0.4%, respectively.

The larger the λ, the more weight is given to the current observation and more quickly the weight tails off over previous observations. Larger values of λ therefore produce a less smooth EWMA trace, because each new observation changes the EWMA statistic by a larger amount. Larger values of λ also make the EWMA trace more responsive to change, moving back inside the limits more quickly after an alarm has been triggered and performance has returned to the baseline rate.

For chart comparisons, a λ of 0.01 is selected and the control limits are adjusted to give the same in-control ARL as the Shewhart p-chart.

The CUSUM chart

We have used a version of the CUSUM for individual binary data.[22] It is set up to detect a shift from the baseline rate to a specified alternative rate, with the shift expressed as an OR. For chart comparisons, the alternative rate corresponds to an OR of 2 to detect increases, and 0.5 to detect decreases. In general, ORs should be selected to correspond to the smallest change that needs to be detected by monitoring. The value for the control limit was then chosen to give the same in-control ARL as the p-chart.

RESULTS OF CHART COMPARISONS

For small absolute increases in rates of less than 10%, the CUSUM detected change most quickly, followed by the EWMA and then the Shewhart p-chart (table 2). For example, for a baseline rate of 1%, the Shewhart p-chart detected a 2% absolute increase in the rate after 600 patients or procedures on average; the EWMA after 350; and the CUSUM after 320.

Larger absolute increases (>10%) were detected more quickly on average, and the Shewhart p-chart and EWMA chart then had similar out-of-control ARLs for the same in-control ARL. The CUSUM was more efficient even for larger increases. In our comparisons, the g-chart was unable to detect increases in rates (decreases in observations between events) because the lower control limit was at or close to zero.

The CUSUM with optimal settings outperformed the other charts for detecting decreases in rates. In contrast to the p-chart and EWMA, the g-chart was able to detect decreases in low event rates. The other charts could not detect decreases because the lower control limits were at zero. To be able to detect decreases in low event rates, control limits would need to be narrowed, or volumes per period would need to be increased for the p-chart, and the weight λ reduced for the EWMA.

EFFECTS OF CHART SETTINGS

To describe effects of altering settings for the p-chart, EWMA and CUSUM, we refer to results presented in table 3 and tables A1 and A2 in the web supplementary appendix. We illustrate the impact of different settings using our example of monitoring mortality after bowel cancer surgery.

Table 2 Comparison of out-of-control ARLs, in number of observations, for the Shewhart p-chart*, the EWMA† the CUSUM‡ and the g-chart

		Size of Shift		Out-of-control ARL			
Baseline event rate	In-control ARL	Absolute shift	OR	p-chart	EWMA	CUSUM	g-chart
1%	14 400	−0.5%	0.5	N/A§	N/A§	1 500	2 400
		+1%	2.0	1 000	990	750	N/A§
		+2%	3.1	600	350	320	N/A§
10%	14 600**	−5%	0.5	500	520	270	690
	9 800	+5%	1.6	400	400	330	N/A§
	9 800	+10%	2.3	200	170	70	N/A§
50%	8 750	−15%	0.5	200	250	110	500
		+15%	1.9	200	250	110	N/A§
		+25%	3.0	100	160	50	N/A§

*3-sigma limits and supplementary runs rules. Reporting periods=200, 100 and 50 cases for 1%, 10% and 50% baseline event rates.
†Value of λ=0.01.
‡OR=2 to detect increases and ½ to detect decreases in rates.
§The lower-limit of the chart is at zero for the selected in-control ARL and chart settings, such that decreases in rates cannot be detected by the p-chart or EWMA, and increases in rates cannot be detected by the g-chart.
**In-control ARLs for a chart set up to detect a decrease can differ from those for a chart set up to detect an increase because the exact binomial probability of crossing the upper limit will not always be the same as the exact binomial probability of crossing the lower limit.
ARL, average run length; CUSUM, cumulative sum; EWMA, exponentially weighted moving average.

For the g-chart, in its most common version, only the control limits can be altered to achieve the desired probability of a false alarm or in-control ARL.

Different reporting periods for the Shewhart p-chart

With fixed 3-sigma control limits, lengthening the reporting period lengthens the in-control ARL and the out-of-control ARL. These can become very long, leading to risk-averse charts that have a very low false alarm rate, but which take a long time to detect a genuine change in performance (table 3). Conversely, shortening periods reduces the in-control ARL and can lead to a high rate of false alarms.

As the reporting period is lengthened, the in-control ARL increases more than the out-of-control ARL. For example, for a baseline rate of 1%, increasing the volume per reporting period from 30 to 200 increases the in-control ARL from 840 to 14 400. The corresponding out-of-control ARLs for detecting a 2% increase are 150 and 600 (table 3). This improves chart efficiency, meaning that the relative chance of a false alarm versus a genuine signal decreases.

The ARLs do not increase smoothly with the number of observations per period because probability distributions for binary data are discrete (table 3).

Revisiting the bowel surgery data in figure 1, a Shewhart p-chart with quarterly reporting instead of monthly reporting would signal slightly earlier, with two consecutive points above the 2-sigma limits by 12 months.

Table 3 Comparison of out-of-control ARLs, in number of observations, for Shewhart p-chart with 3-sigma limits with different reporting periods

Volume per period:			30		50		100		200		500	
Baseline event rate	Absolute shift	OR	In-control ARL	Out-of-control ARL	In-control ARL	Out-of-control ARL	In-control ARL	Out-of-control ARL	In-control ARL	Out-of-control ARL	In-control ARL	Out-of-control ARL
1%	−0.5%	0.5	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	N/A*	61 000	5 000
	+1%	2.0	840	270	2 150	400	3 200	600	14 400	1 000	48 500	1 500
	+2%	3.1	840	150	2 150	200	3 200	300	14 400	600	48 500	1 000
10%	−5%	0.5	2 700	330	6 150†	500	14 600†	500	38 800†	600	1 15 000†	1 000
	+5%	1.6	2190	270	5 850	400	9 800	400	29 000	600	94 500	1 000
	+10%	2.3	2 190	120	5 850	150	9 800	200	29 000	400	94 500	1 000
50%	−15%	0.5	4 710	180	8 750	200	16 000	200	43 000	400	1 09 500	1 000
	+15%	1.9	4 710	180	8 750	200	16 000	200	43 000	400	1 09 500	1 000
	+25%	3.0	4 710	60	8 750	100	16 000	200	43 000	400	1 09 500	500

*Lower-limits of chart are at zero so decreases cannot be detected.
†In-control ARLs for a chart set up to detect a decrease can differ from those for a chart set up to detect an increase because the exact binomial probability of crossing the upper limit will not always be the same as the exact binomial probability of crossing the lower limit.
ARL, average run length.

Different weights for the EWMA

For a fixed in-control ARL, values of the weight λ can be selected to minimise the out-of-control ARL and maximise chart efficiency. Control limits are simultaneously adjusted to achieve the desired in-control ARL.

The optimal value of λ depends on the baseline rate and the shift size to be detected, with higher baseline rates having a larger optimal λ. Weights of around 0.01 perform well across a range of rates (online supplementary table A1).

In figure 1 the EWMA was set up with a λ of 0.01, and triggers an alarm in month 12. If a λ of 0.005 is used instead, the EWMA with the same in-control ARL signals slightly earlier, in month 10 (approximately 400th procedure).

Different settings for the CUSUM

The CUSUM chart is faster at detecting increases in rates than the other charts, even when the shift that it is designed to detect differs from the actual shift (online supplementary table A2). It is fastest when the hypothesised shift matches the actual shift.

An OR of 2 corresponds to a realistic increase in rate across a range of baseline event rates (online supplementary table A2) and an OR of ½ corresponds to a realistic decrease.

The CUSUM in figure 1 was set up to have an alternative rate of 5% (OR of 1.7) and triggers an alarm in month 9. It triggers in the same month if an OR of 2 is used instead.

DISCUSSION

Main findings

For small absolute increases in event rates, less than around 10%, the CUSUM detects changes more rapidly than the EWMA, and the EWMA detects changes more rapidly than the Shewhart p-chart.

For larger increases in rates, over 10%, the EWMA and Shewhart p-chart both have good performance, although not quite as good as the CUSUM.

The g-chart can detect decreases in low event rates, whereas the p-chart may not be able to do so without increasing the false signal rate substantially. The CUSUM with optimal settings is the most efficient for detecting decreases.

The Shewhart p-chart is the most accessible chart, particularly in terms of setting up the chart. However, the choice of reporting period can have a large effect on its behaviour. In contrast, the CUSUM is the most difficult of the four charts to construct and interpret.

Strengths and limitations of our study

There are many charts available for monitoring binary data and we have not covered all of them. To take one example, a different EWMA chart for rare events has been proposed that, like the g-chart, monitors the number of cases between events.[27 28]

There are also different versions of the charts that we have reviewed that could affect comparisons of their performance. For example, we used the Shewhart p-chart with 3-sigma limits and supplementary runs rules based on the 2-sigma limits. Using runs rules has previously been shown to be more efficient for detecting shifts than narrowing the limits of the basic Shewhart chart when monitoring means of continuous variables.[24] We compared the performance of the Shewhart p-chart with adjustable control limits to the other charts, and our conclusions were unaltered.

Our comparisons cover a range of changes in rates that were selected to be realistic for binary clinical outcome or process-of-care data. The range does not cover very large changes. However, very large changes should be clear in the raw data. Persistent but smaller changes are arguably more important to detect since they may otherwise go unnoticed or disputed without statistical methods for detecting them.[3]

We used ARLs to compare the charts. Probability distributions of run lengths are highly skewed, so that the median run length is often well below the mean, so that more than half of false alarms would occur before the designed ARL. Another option would be to design charts with specified probabilities of detecting a change and signalling a false alarm over a given time frame for monitoring.[3] Methods have also been proposed for directly controlling the conditional probability of a false alarm.[29]

We have presented comparisons for charts set up to have an in-control ARL of 10 000 patients or procedures. However, the comparative efficiency of the charts was similar when they were designed with shorter in-control ARLs. For example, with standard 3-sigma limits, the g-chart set up to detect a decrease in rate from 1% to 0.5% had an in-control ARL of 5400 and an out-of-control ARL of 1460. The CUSUM with the same in-control ARL had an out-of-control ARL of 1000.

We have also made several assumptions that may influence chart behaviour and our conclusions about their comparative efficiency. We have assumed that the baseline event rate is known. In practice, data collection may not precede the start of monitoring, or a very low event rate may be poorly estimated even when prior data are collected. Charts based on numbers between events have been shown to be more robust than the CUSUM to poor estimation of a low event rate.[30] Technical methods have been proposed to adjust for uncertainty due to parameter estimation.[31] In clinical applications, an external target rate may alternatively be used, although this would alter interpretation of the chart and any resulting signals.

We have assumed that shifts in performance occur at the start of the monitoring period. The Shewhart p-chart and g-chart do not update with each observation. If a shift in performance occurs partway through a reporting period for the Shewhart p-chart, or in

between two events for the g-chart, this will increase their out-of-control ARLs. Simulations of shifts in the performance partway through monitoring had little effect on our estimates of ARL (results not shown) and did not change the findings of the study.

We have assumed that the probability of an event is constant over time under the null hypothesis that clinical performance is unchanged. In practice, clinical outcome data are often prone to systematic variability other than changes in performance such as seasonality, variation in patient risk factors and changes in data quality. Based on simulated data, in-control ARLs have been shown to be shorter when monitoring outcomes in high-risk groups of patients, leading to a higher rate of false alarms than the chart is designed with.[32] However, comparisons of risk-adjusted versions of the charts have found the CUSUM to be more efficient across simulated variations in case mix[16] and the risk-adjusted CUSUM has been recommended for monitoring surgical outcomes.[4 21 33]

Comparison with the literature

Studies examining the ARL performance of different charts are standard in the industrial SPC literature and not as common in the healthcare SPC literature. For binary data, other studies agree that the CUSUM and the EWMA are quicker than the Shewhart p-chart for detecting small increases in rates, and several reviews have recommended the CUSUM for individual binary data, including surgical-outcome data.[4 16 18]

Differences between industrial and healthcare settings limit the applicability of some of the findings from the former to the latter. For example, binary outcome data are more common in healthcare applications. Comparisons in the industrial SPC literature demonstrating better performance of the Shewhart p-chart for detecting large shifts do not always apply to monitoring individual binary data. Shifts of 2 or 3 SDs are often defined as large,[15] but such large shifts are rare with binary clinical data. This difference in the definition of small and large shifts may explain the finding of Grigg and Farewell that the Shewhart p-chart performs worse than the CUSUM even for 'large' shifts in rates. Here they characterise a large shift using an OR of 5. This corresponds to an increase from 6.6% to 26.1% in their example of mortality after cardiac surgery, but is a shift of just 1 SD in the individual data.[16] A comparison of the CUSUM and EWMA that found them to have similar performance was for monitoring a continuous variable, not a binary variable.[34]

Monitoring in healthcare typically covers 100% of patients so that use of the Shewhart p-chart requires data to be aggregated over a reporting period such as a week or a month. In contrast, sampling the output of a process at short intervals is common in industrial applications, and sample size is often not a constraint.[7] The effect of data aggregation on the efficiency of the CUSUM chart was examined by Reynolds and Stoumbos, who compared the CUSUM for individual binary data and for grouped binary data.[22] They found that the former has shorter out-of-control ARLs than the latter for a fixed in-control ARL, although both charts performed better than the Shewhart p-chart.

Related to this, the Shewhart p-chart can't always be optimised in healthcare applications in the same way that it can for many industrial applications. Where sample sizes are not a constraint, the rule of thumb of $np \geq 5$ observations is often used, where n is the sample size per period and p is the event rate. Sample sizes larger than this are recommended for detecting reductions in rates.[35] The rule is based on the threshold at which the binomial distribution can be approximated with the normal distribution and 3-sigma limits perform as intended.[22] In healthcare applications, it can be difficult to achieve sufficiently large sample sizes without aggregating data over very long reporting periods, potentially causing delays in detection of changes until the end of the reporting period.[16] Another option would be to use exact binomial limits in the place of 3-sigma limits.

Practical implications

We have compared the efficiency of charts for detecting changes in rates of binary data under ideal conditions. However, other practical considerations will influence the choice of chart.

One consideration is the frequency of the procedure or prevalence of the condition for which outcomes or processes-of-care are being monitored. For low volume procedures and rare conditions, it may not be possible to achieve volumes per reporting period that achieve desired ARLs. For example, oesophagogastric cancer surgery is performed on 25–30 patients per year in a typical English hospital. With postoperative mortality as low as 3%, we have suggested 200 patients per reporting period to give an in-control ARL of 10,000 patients. For this procedure, reporting periods of several years would be needed, which is inappropriate for continuous monitoring. On the other hand, shorter in-control ARLs may be more appropriate, since even short ARLs correspond to long periods of calendar time. In such settings, it is useful to estimate the in-control ARL when selecting reporting periods, using table 3 as a guide. There is also a stronger case for using the CUSUM or EWMA that can update with each procedure.

The practical importance of false alarms and detection times will also vary. For a quality improvement team monitoring a process of care such as timely assessment or prescription of an appropriate medication, false alarms are not necessarily such a problem as for monitoring of outcomes. Each signal can be investigated in detail and action taken without compromising patient safety or causing unnecessary work. In this situation, the Shewhart p-chart with runs rules

may represent the best choice and shorter reporting periods, corresponding to shorter in-control ARLs, may be appropriate. The g-chart may also be useful for detection of improvement. However, we still think it is important that users are aware of the impacts of their choices on ARLs and again can use tables 2 and 3 as a guide.

In contrast, false alarms may be more detrimental when reporting is open to the public rather than used solely for internal performance review, particularly when reporting outcomes such as mortality.[36] For patient safety indicators, early detection of a problem is also important. In these cases, charts should be designed with sufficiently long in-control ARLs and the efficiency of the CUSUM and EWMA may be a particular advantage. We have suggested settings for these charts and also provide the in-control ARLs for different control limits in the online supplementary appendix.

Simplicity and ease of interpretation of charts for healthcare teams are important. Both the Shewhart p-chart and the EWMA provide a trace that can be interpreted as an estimate of the current event rate and graphed on a natural scale as a proportion or percentage. The g-chart also has a natural interpretation. A range of tools and how-to videos are also available to help set up the Shewhart charts, such as the Institute for Health Improvement's Open School resources. [http://www.ihi.org/education/IHIOpenSchool/resources/Pages/AudioandVideo/Whiteboard13.asp]

The CUSUM lacks this interpretability. One suggestion is that the CUSUM control limits can be superimposed on an O-E or *Variable Life Adjustment Display* (VLAD) chart that displays the cumulative observed numbers of events (O) minus the expected number (E).[37] There is a need for tools to make these methods more accessible if they are to become more widely used, and also a need for better access to skilled analysts with health services.[38]

Acknowledgements The authors particularly thank the editor, Kaveh Shojania, and the associate editor, Tom Woodcock, for their thoughtful revisions and substantial contribution to the thinking in this paper. The authors also thank their reviewers for useful comments and Chris Rogers, Martin Bardsley and Naomi Wakeman for early discussions and suggestions. Hospital episode statistics were made available by the NHS Digital.

Contributors JN and KW are joint first authors. They conceived the idea for the study, conducted the literature search, and synthesised the findings. JN, KW, JvdM and DAC developed the idea for the study. DAC advised on the literature search. JN and KW carried out the analyses. JN and KW drafted and revised the paper. All authors contributed to revising the paper and all reviewed and approved the final draft.

Funding JN was funded by a National Institute for Health Research Postdoctoral Fellowship (PDF-2013-06-078). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

Competing interests None declared.

Patient consent The study involves secondary analysis of anonymised data.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Benning A, Ghaleb M, Suokas A, *et al*. Large scale organisational intervention to improve patient safety in four UK hospitals: mixed method evaluation. *BMJ* 2011;342:d195.
2 Nicolay CR, Purkayastha S, Greenhalgh A, *et al*. Systematic review of the application of quality improvement methodologies from the manufacturing industry to surgical healthcare. *Br J Surg* 2012;99:324–35.
3 Sherlaw-Johnson C, Bardsley M. Monitoring change in health care through statistical process control methods. *The Nuffield Trust* 2016 http://www.nuffieldtrust.org.uk/publications/monitoring-change-health-care-through-statistical-process-control-methods (accessed 22 Dec 2016).
4 Woodall WH, Fogel SL, Steiner SH. The monitoring and improvement of surgical-outcome quality. *J Qual Technology* 2015;47:383–99.
5 Schmidtke KA, Watson DG, Vlaev I. The use of control charts by laypeople and hospital decision-makers for guiding decision making. *Q J Exp Psychol* 2017;70:1114–28.
6 Anhøj J, Hellesøe AB. The problem with red, amber, green: the need to avoid distraction by random variation in organisational performance measures. *BMJ Qual Saf* 2017;26:81–4.
7 Woodall WH. The use of control charts in health-care and public health surveillance. *J Qual Technology* 2006;38:89–104.
8 Schmidtke KA, Poots AJ, Carpio J, *et al*. Considering chance in quality and safety performance measures: an analysis of performance reports by boards in English NHS trusts. *BMJ Qual Saf* 2017;26:61–9.
9 Thor J, Lundberg J, Ask J, *et al*. Application of statistical process control in healthcare improvement: systematic review. *Qual Saf Health Care* 2007;16:387–99.
10 Spiegelhalter DJ. Monitoring clinical performance: a commentary. *J Thorac Cardiovasc Surg* 2004;128:820–2.
11 Aylin P, Best N, Bottle A, *et al*. Following Shipman: a pilot system for monitoring mortality rates in primary care. *Lancet* 2003;362:485–91.
12 Rogers CA, Ganesh JS, Banner NR, *et al*. Cumulative risk adjusted monitoring of 30-day mortality after cardiothoracic transplantation: UK experience. *Eur J Cardiothorac Surg* 2005;27:1022–9.
13 Mohammed MA, Worthington P, Woodall WH. Plotting basic control charts: tutorial notes for healthcare practitioners. *Qual Saf Health Care* 2008;17:137–45.
14 Benneyan JC. Number-between g-type statistical quality control charts for monitoring adverse events. *Health Care Manag Sci* 2001;4:305–18.

15 Montgomery DC. *Introduction to Statistical Quality Control*. 6th ed. New York: Wiley & Sons, 2009.
16 Grigg O, Farewell V. An overview of risk-adjusted charts. *J R Stat Soc Ser A Stat Soc* 2004;167:523–39.
17 Spiegelhalter D, Sherlaw-Johnson C, Bardsley M, *et al*. Statistical methods for healthcare regulation: rating, screening and surveillance. *J R Stat Soc Ser A Stat Soc* 2012;175:1–47.
18 Szarka JL, Woodall WH. A review and perspective on surveillance of Bernoulli processes. *Qual Saf Health Care* 2011;27:735–52.
19 Cook DA, Coory M, Webster RA. Exponentially weighted moving average charts to compare observed and expected values for monitoring risk-adjusted hospital indicators. *BMJ Qual Saf* 2011;20:469–74.
20 Page ES. Control charts with warning lines. *Biometrika* 1955;1:243–57.
21 Steiner SH, Cook RJ, Farewell VT. Risk-adjusted monitoring of binary surgical outcomes. *Med Decis Making* 2001;21:163–9.
22 ReynoldsMR Jr, Stoumbos ZG. A CUSUM chart for monitoring a proportion when inspecting continuously. *Journal of Quality Technology* 1999;3:1.
23 Hospital Episode Statistics. http://www.hscic.gov.uk/hes (accessed Mar 2016).
24 Champ CW, Woodall WH. Exact results for Shewhart Control Charts with Supplementary runs rules. *Technometrics* 1987;29:393–9.
25 Jc F, Spiring FA, Xie H. On the average run lengths of quality control schemes using a Markov chain approach. *Statistics & Probability Letters* 2002;56:369–80.
26 Bland JM. The odds ratio. *BMJ* 2000;320:1468.
27 Spliid H. Monitoring medical procedures by exponential smoothing. *Stat Med* 2007;26:124–38.
28 Spliid H. An exponentially weighted moving average control chart for Bernoulli data. *Qual Reliab Eng Int* 2010;26:97–113.
29 Gombay E, Hussein AA, Steiner SH. Monitoring binary outcomes using risk-adjusted charts: a comparative study. *Stat Med* 2011;30:2815–26.
30 Lee J, Wang N, Xu L, *et al*. The Effect of Parameter Estimation on Upper-sided Bernoulli Cumulative Sum Charts. *Qual Reliab Eng Int* 2013;29:639–51.
31 Weiss CH, Atzmüller M. EWMA control charts for monitoring binary processes with applications to medical diagnosis data. *Qual Reliab Eng Int* 2010;26:795–805.
32 Tian W, Sun H, Zhang X, *et al*. The impact of varying patient populations on the in-control performance of the risk-adjusted CUSUM chart. *Int J Qual Health Care* 2015;27:31–6.
33 Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Methods Med Res* 2003;12:147–70.
34 Lucas JM, Saccucci MS. Exponentially Weighted moving average control schemes: properties and Enhancements. *Technometrics* 1990;32:1–12.
35 Morris RL, Riddle EJ. Determination of Sample size to detect Quality Improvement in p-Charts. *Qual Eng* 2008;20:281–6.
36 Lilford R, Mohammed MA, Spiegelhalter D, *et al*. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;363:1147–54.
37 Sherlaw-Johnson C. A method for detecting runs of good and bad clinical outcomes on Variable Life-Adjusted Display (VLAD) charts. *Health Care Manag Sci* 2005;8:61–5.
38 Bardsley M. Understanding analytical capability in health care. *Do we have more data than insight*. London: The Health Foundation, 2016. http://www.health.org.uk/sites/health/files/UnderstandingAnalyticalCapabilityInHealthCare.pdf. (accessed 22 Dec 2016).