International Journal for Quality in Health Care 2008; **Volume 20, Number 1**: pp. 31–39 10.1093/intqhc/mzm060
Advance Access Publication: 8 December 2007

Using control charts to monitor quality of hospital care with administrative data

MICHAEL COORY[1,2], STEPHEN DUCKETT[1,2] AND KIRSTINE SKETCHER-BAKER[2]

[1]School of Population Health, University of Queensland, Australia, and [2]Reform and Development Division, Queensland Health, Australia

Abstract

Objective. Nearly all hospital-specific comparative analyses, based on administrative data, are presented using cross-sectional displays. In this paper, we compare cross-sectional analyses with sequential monitoring using control charts.

Design. Analysis of administrative data to compare cross-sectional funnel plots with one type of control chart: the risk-adjusted, expected-minus-observed plot.

Setting. Eighteen tertiary and base hospitals in Queensland, Australia, for the two financial years 2003–04 and 2004–05.

Participants. Patients admitted with acute myocardial infarction.

Main outcome measure. Risk-adjusted, 30-day, in-hospital, mortality rates.

Results. There were no outliers on the cross-sectional funnel plots for either of the 2 years using three-sigma limits and three low-outliers and one high-outlier using two-sigma limits. One reasonable interpretation of these plots is that most of the variations are due to statistical noise and there is little to be learnt by seeking to understand the reasons for variation across hospitals. In contrast, for the control charts, 28% of hospitals signalled for a relative increase of 75% above that for all hospitals combined.

Conclusion. If the aim of clinical indicators based on administrative data is to provide a starting point for learning, then control charting provides potentially more useful information than the more commonly used cross-sectional analyses. Control charts provide an understandable and up-to-date overview that allows early detection of runs of good or bad outcomes that can help hospitals identify areas for more in-depth self-monitoring and learning.

Keywords: control charting, data interpretation, information dissemination, quality indicators

By definition, hospital administrative data are primarily collected for funding and other administrative purposes, not for assessing quality-of-care. However, such data are increasingly being used to derive hospital-specific clinical indicators. The secondary nature of these analyses means that there are particular concerns about residual confounding and measurement error, with several commentators arguing that such analyses should not be used to make definitive judgements about performance. Instead they should be used to help hospitals identify areas for more in-depth self-monitoring [1–3].

In spite of the limitations, it is likely that funding, purchasing or coordinating agencies will continue to conduct hospital-specific, comparative analyses based on administrative data. There are a variety of reasons for this, some associated with value judgements about the need for accountability and the lack of alternative data [4], and others related to the perceived failure of hospital-based audit processes [5]. The main issues now are not so much about whether such data should be used, but about how to present them in a way that conveys the most information [4].

Nearly all programs that use administrative data for hospital-specific comparative analyses present cross-sectional analyses [6]. That is, risk-adjusted hospital-specific outcomes (e.g. mortality, length-of-stay) for particular conditions (e.g. acute myocardial infarction, pneumonia) or procedures (e.g. cardiac bypass surgery) are aggregated over a set period, often 12 months. This paper compares cross-sectional analyses with sequential monitoring using control charts. The aim is to illustrate the characteristics of control charts, which are particularly good at detecting short runs of unusual outcomes and providing opportunities for learning that might be missed in cross-sectional analyses.

Methods

Data

Data were obtained from the Queensland Hospital Admitted Patients Data Collection (QHAPDC), which contains, *inter*

[1]Address reprint requests to: Michael Coory; E-mail: m.coory@uq.edu.au

alia, the demographic characteristics of the patients, the principal diagnosis, other conditions treated and the procedures performed. Queensland is the north-eastern state of Australia with a population of 4.0 million, or 18% of the total Australian population. QHAPDC is similar to administrative hospital databases in the other states and territories of Australia and the USA, the UK and Canada.

We chose the 30-day, in-hospital mortality rate following admission for acute myocardial infarction to conduct this comparison of methods because it is a commonly used clinical indicator derived from administrative data [7]. Patients with acute myocardial infarction were identified using the 'International Classification of Diseases' International Classification of Diseasesversion 10 codes I21*x*–I22*x* [8]. The data analysed were from 18 tertiary and base hospitals in Queensland, for the two financial years 2003–04 and 2004–05.

In the Queensland Quality Measurement Program, the clinical indicator for acute myocardial infarction has three inclusion criteria: admitted through the emergency department of the hospital, age between 30 and 89 years, and died or discharge status of alive with length-of-stay greater than 3 days. The aim of these criteria is to reduce the number of false-positive diagnoses of acute myocardial infarction in administrative data [9].

Risk-adjusted, expected-minus-observed plots

Clinical indicators should be risk-adjusted for potential confounders (to the extent that this is possible using administrative data) so that hospitals that treat higher risk patients are not unfairly penalized. There are several types of risk-adjusted control charts suitable for use in health applications [10]. For this comparison, we chose the risk-adjusted, expected-minus-observed plot. These are also called risk-adjusted cumulative sum (CUSUM) plots [11], Cumulative Risk-Adjusted Mortality (CRAM) plots [12], Variable Life Adjusted Display (VLAD) charts [13] and cumulative excess mortality charts [14]. We chose this type of control chart because it gives an intuitive display of net number of 'actual' outcomes (e.g. deaths) versus the number 'expected' and is routinely used in some clinical departments [15].

The risk-adjusted, expected-minus-observed plot has the form:

$$C_n = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i$$

(i) X_i is the *actual* outcome for the *i*th patient; for this analysis, 1 if the patient died and 0 if the patient survived.

(ii) Y_i is the 'expected' risk (i.e. it takes values between 0 and 1) of the outcome (e.g. death) for the *i*th patient assuming he or she has the same risk as the average for all hospitals combined for his or her age, sex and co-morbidity pattern. For this analysis, Y_i was calculated from a logistic regression model, which included five-year age groups as indicator variables, sex and co-morbidities. The co-morbidities were shock, dysrhythmias, congestive heart failure, hypertension, diabetes, chronic renal failure, dementia, stroke and malignancy. These co-morbidities were identified in other studies [9, 16] (and confirmed in our data) as predicting short-term mortality for acute myocardial infarction. The plots are updated each month and the risk adjustment model is re-calibrated each month using the latest 12 months of data.

The position on the plot represents the number of outcomes (e.g. deaths) at that hospital subtracted from the number expected if that hospital had the same distribution of age, sex and co-morbidities as all hospitals combined. A plot that stays near the horizontal zero line indicates that the number of deaths is similar to that expected. If a point on the plot is below the horizontal zero line, then the number of deaths at that time was greater than expected; if the point is above the horizontal line, then the number of deaths is fewer than expected. A run of more deaths than expected is represented by a downward trend and a run of fewer deaths than expected is represented by an upward trend [15].

Thresholds

The traditional way of describing the statistical characteristics of a control chart is in terms of average run length to false alarm and average run length to true alarm because the familiar Type 1 and Type 2 error rates are not appropriate. With sequential monitoring the Type 1 error rate is not constant, but increases with the length of the monitoring period. The probability of eventually signalling an alarm is 1.0 for all sequential tests, so that the Type 1 error rate will eventually be 100% [17].

Ideally, average run length to false alarm should be long (analogous to a low Type 1 error rate) and the average run length to true alarm should be short (analogous to high statistical power). In practice, there is a trade-off and a good choice for a threshold, where the chart is said to signal, is one where the average run length to true alarm is suitably short and the average run length to false alarm is not unacceptably short.

Average run lengths for different thresholds can be estimated using simulations, Markov chains, or in certain circumstances by approximating formulae [18]. In this paper, we used simulations. Briefly, we specified data sets of 10 000 patients (under the null and various alternative hypotheses) and iterated 10 000 times to obtain estimates of the median ARL to true and false alarm.

Alternative hypotheses were pre-specified as relative risk reductions of 30, 50 or 75% or relative risk increases of 30, 50 or 75%. These were converted to odds ratios for use in the simulations to calculate thresholds (usually denoted as *h*). As others have done, we used the log-likelihood-ratio form of the CUSUM to obtain the thresholds because of it is more mathematically convenient than the expected-minus-observed plot [15]. A description of the log-likelihood-ratio CUSUM is given by Grigg *et al.* [11].

Table 1 Thresholds (h) and average run length to false alarm for an average run length to true alarm of 100 and pre-specified relative risk reductions and relative risk increases

	Relative risk reduction			Relative risk increase		
	30%	50%	75%	30%	50%	75%
h	2.6	3.6	4.9	2.8	3.7	5.0
Average run length to false alarm	229	682	2447	264	834	3118

Having thresholds for relative risk reduction as well as relative risk increase emphasizes that we want to learn from runs of both good and bad outcomes. Having three thresholds for each emphasizes that the plot is not intended to make definitive yes/no judgements, rather it is to be used to identify when local investigation is warranted.

Control charts should be linked to specific learning actions [19]. The plan in Queensland is to implement a system such that if the chart signals for a relative risk increase of 30%, then the hospital would be advised to investigate. If the chart signals at a relative risk increase of 50%, then the Area Health Service would be advised to investigate. If the chart signals at 75%, then the Patient Safety and Quality Board would be notified.

Table 1 shows the results of the simulations to obtain thresholds (h) for the pre-specified relative risk reductions (and increases) of 30, 50 and 75% and a pre-specified average run length to true alarm of 100 patients. After obtaining these thresholds (Table 1, row 1) we then used them in subsequent simulations to obtain average run lengths to false alarm (row 2).

Starting and reset values

In industry, the starting value for a control chart is usually set at zero because the machine or process is known to be in control (i.e. the machine has been recently calibrated) at the start of monitoring. This is not usually the case in health applications, where, before monitoring starts, there is no reason to believe the process is either in-control or out-of-control. We therefore set the initial log-likelihood-ratio CUSUM value at $h/2$; that is, half the threshold value. As Grigg *et al.* state, this makes sense intuitively because it reflects uncertainty as to whether the process is in-control at the start of monitoring. It avoids the problem of missing early runs of poor outcomes [11].

Similar comments apply to resetting: in industrial application, the machine or process is monitored until it is out-of-control. Monitoring then stops, the machine is recalibrated, and therefore known to be in-control, and monitoring restarts with a log-likelihood-ratio CUSUM value of zero. In health applications, there is usually no reason to believe the process will be in-control after a signal and, as with the starting value, it is probably better to start the monitoring again at $h/2$, rather than zero.

Cross-sectional analyses

The aim of this paper is to compare cross-sectional analyses and sequential monitoring using control charts. A common method of cross-sectional analysis is to calculate the risk-adjusted mortality ratio for each hospital over a set period, say 12 months. Using the previous notation, the risk-adjusted mortality ratio can be written as $\sum X_n/\sum Y_n$. To communicate this information, one option is the caterpillar plot where three-sigma limits (equivalent to 99.8% confidence intervals) are plotted for each hospital. If the three-sigma limits for a particular hospital do not include the average for all hospitals combined (1.0), then the hospital is flagged as an outlier and suitable for further investigation. An outlying hospital can either be a high-outlier (risk-adjusted mortality rates greater than the average for all hospitals combined), or a low-outlier (risk-adjusted mortality rates less than the average for all hospitals combined). Another option is to use two-sigma limits (equivalent to 95% confidence intervals), but these carry with them the concern that too many hospitals might be flagged as outliers because of the problem of multiple statistical comparisons.

One problem with caterpillar plots is that they lead the user to focus on a spurious rank ordering of hospitals [20]. A better way of presenting such data is the funnel plot in which hospital-specific measures are plotted against their statistical precision, so that the confidence limits funnel around the average for all hospitals combined (1.0). For this paper, we used funnel plots with two- and three-sigma limits as defined by formulae given by Spieleghalter [21].

Results

After applying the inclusion criteria, there were 4158 admissions for acute myocardial infarction to the 18 tertiary and base hospitals in Queensland for the 2-year study period (2079 admissions per year). The median number of admissions per hospital, per year was 103; range (40–265); interquartile range (74–154). The average 30-day, in-hospital, mortality rate for all hospitals combined was 12.4%.

On the basis of the three-sigma criterion, no hospital flagged in either financial year, either as a high or low-outlier (Fig. 1). In 2003–04, the three hospitals that were low-outliers on the two-sigma criterion might have been scrutinized further, so that all hospitals could potentially learn from their lower than average rates. The hospital with the highest risk-adjusted mortality rate in 2003–04 (labelled A in Fig. 1) just failed to signal at the two-sigma limit. In 2004–05, one hospital was a high-outlier based on the two-sigma criterion (labelled B in Fig. 1) and further investigation of this hospital might have been useful.

In short, the overall impression from the funnel plots is that most of the hospitals lie within two-sigma limits. In

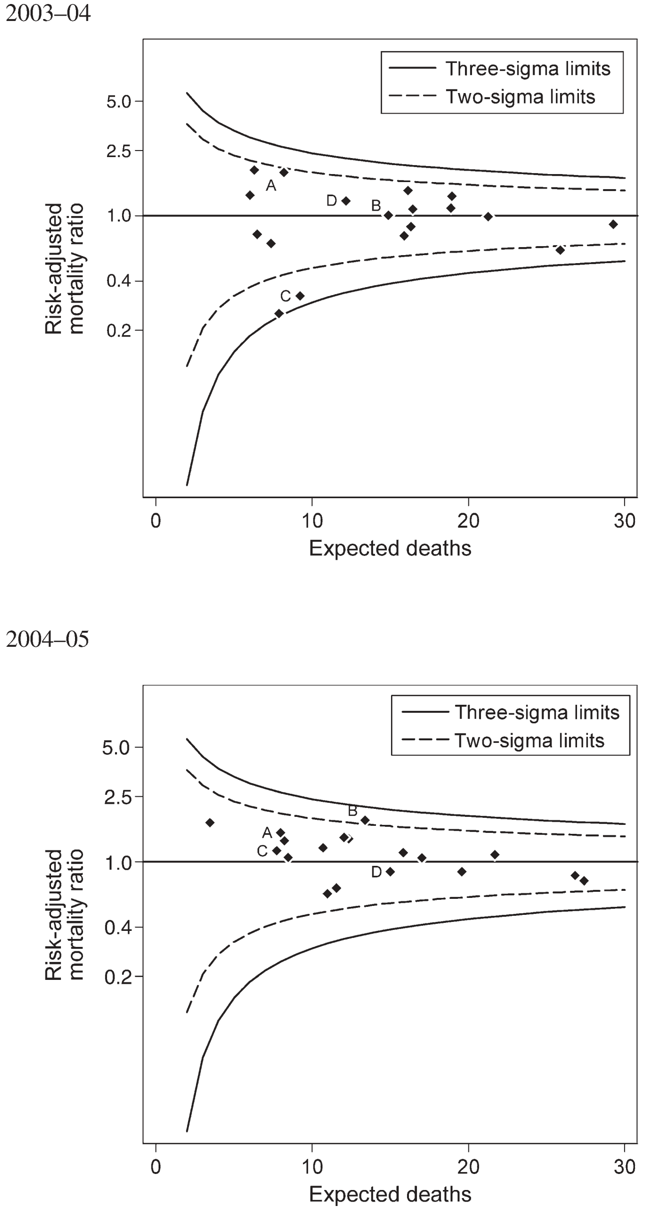

Figure 1 Cross-sectional funnel plots. Labels A–D refer to particular hospitals whose control charts are shown in Figs 2–5.

contrast, there were several signals from the control charts; 5 of the 18 hospitals (28%) flagged for a relative risk increase of 75% (Table 2).

Figs 2–5 show the risk-adjusted, expected-minus-observed plots for four of the hospitals. In September 2003, Hospital A (Fig. 2) signalled once at a relative risk increase of 75% and twice at relative risk of 50 and 30% (signals at 30% not shown if they coincided with the 50% signals). The signals draw attention to the difference between the seven deaths that occurred during August and September 2003 and the 1.7 that were expected.

For the entire financial year 2003–04, Hospital A had 95 admissions for acute myocardial infarction and 15 deaths. The expected number of deaths was 8.2 which gave a risk-adjusted mortality ratio of 1.83 (95% confidence interval: 0.99–3.11). That is, for 2003–04 this hospital just failed to signal at the two-sigma limit on the funnel plot. During 2004–05, there were 91 admissions and the control chart was roughly horizontal, with only one signal at a relative risk increase of 30%. For that financial year, 12 deaths were observed when 8.0 were expected; risk-adjusted mortality ratio: 1.50; 95% confidence interval: 0.78–2.62.

The number of deaths at Hospital B (Fig. 3) was similar to the number expected from July 2003 to August 2004, but for the 8-month period September 2004–April 2005, there were 20 deaths, when only 10 were expected. The hospital signalled at the 30% threshold on 3 November 2004 and 10 March 2005, at the 50% threshold on 7 February 2005 and

Table 2 Signals from control charts for 2 years 2003–04 to 2004–05

	Relative risk reduction			Relative risk increase		
	30%	50%	75%	30%	50%	75%
No. of flags over 2 years	10	7	3	16	8	5
Flags per 1000 admissions	2.4	1.7	0.7	3.8	1.9	1.2
No. of hospitals that flagged at least once by size						
40–74 admissions per year (5 hospitals)	2	2	2	1	0	0
75–99 admissions per year (4 hospitals)	0	0	0	4	4	4
100–149 admissions per year (5 hospitals)	2	1	0	2	1	1
159–265 admissions per year (4 hospitals)	2	2	1	2	1	0
Total no. of hospitals that flagged at least once	6	5	3	9	6	5
% of hospitals that flagged at least once	33	28	17	50	33	28

at the 75% threshold on 23 April 2005. This hospital signalled at the two-sigma level on the funnel plot for financial year 2004–05, but not at the three-sigma level (observed: 24; expected: 13.4; risk-adjusted mortality ratio: 1.79; 95% confidence interval: 1.15–2.67). For the previous financial year 2003–04, when the control chart was flat, the observed number of deaths was 15 and the expected number was 14.9 (risk-adjusted mortality ratio: 1.01). This hospital had 96 admissions for acute myocardial infarction in 2003–04 and 100 in 2004–05.

Fig. 4 shows the plot for one of the hospitals that signalled as a low-outlier at the two-sigma level in the 2003–04 funnel plot. As a final example, the risk-adjusted case-fatality rate for Hospital D (Fig. 5) was about the state average for the entire 2-year period.

Discussion

If the aim of clinical indicators, based on administrative data, is to provide a starting point for learning, then control charting potentially provides more useful information than the more commonly used cross-sectional analyses. Control charts display details of the history of outcomes at a particular hospital and in many cases, learning actions could be instigated based on the plot itself without using thresholds; for example, the abrupt appearance of a downward slope as in Fig. 3.

Lovegrove *et al.* did not use thresholds for their chart on cardiac surgery, arguing that this would imply what is, or is not, acceptable performance [13]. We agree that control charts should not be used to make definitive judgements, but thresholds can be useful for simplifying and standardizing procedures that identify when the data are worth a closer look. Thresholds should not be used to label poor performance, but rather to identify when investigation is warranted.

There were several signals from the control charts, but none from the cross-sectional charts at the three-sigma limits and four at the two-sigma limits. One sensible interpretation of the cross-sectional charts is that most of the variation is

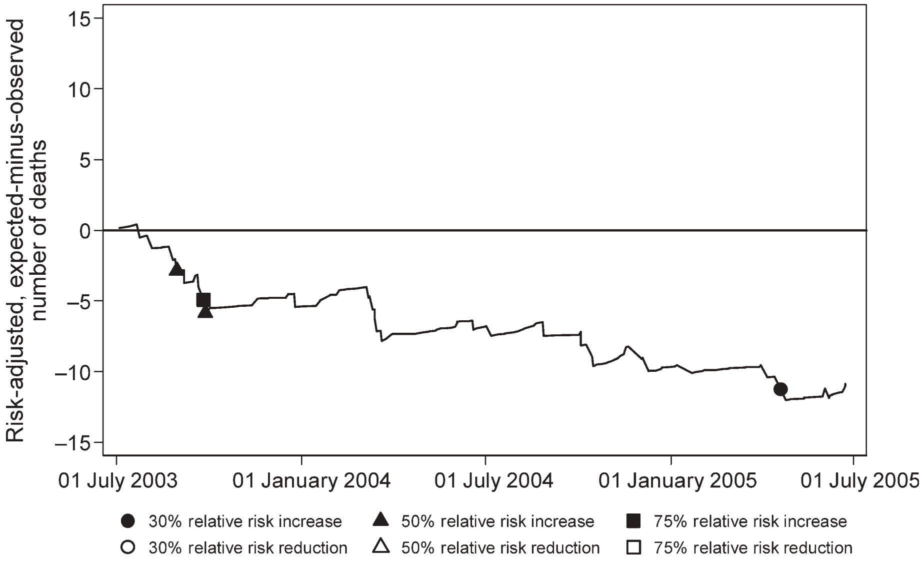

Figure 2 Risk-adjusted, expected-minus-observed plot, Hospital A.

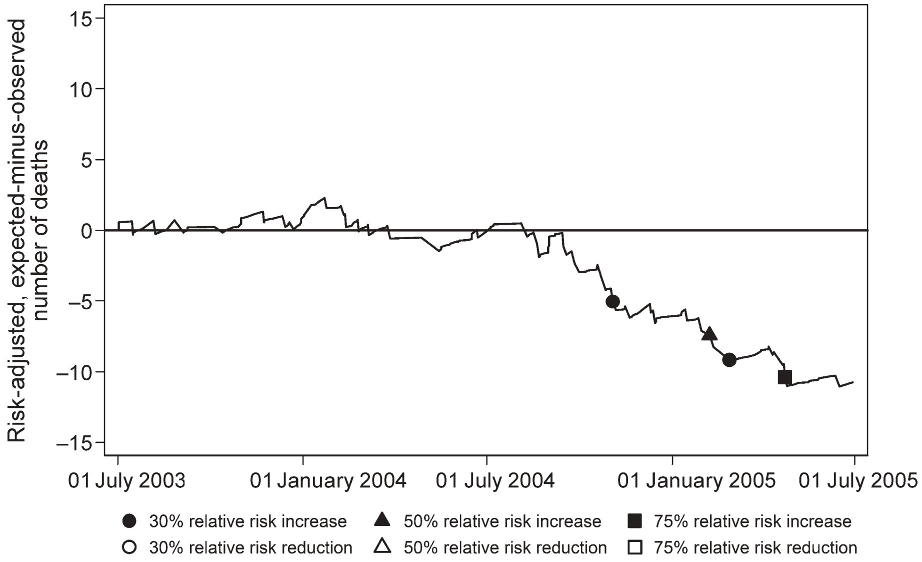

Figure 3 Risk-adjusted, expected-minus-observed plot, Hospital B.

due to statistical noise and there is little to be learnt by seeking to understand the reasons for such variation across hospitals [21].

We do not think that the signals from the control charts in this study are 'statistical' false alarms. The statistical characteristics of the control charts are summarized in Table 1. For a relative risk increase of 75% the average run length to false alarm is 3118 admissions (for a pre-specified average run length to true alarm of 100 admissions). The median number of admissions for acute myocardial infarction per hospital per year was 103; so that for an average hospital a statistical false alarm would occur once every 30 years (3118/103). For a relative risk increase of 30%, a statistical false alarm would occur about once every 2 or 3 years (average run length to false alarm = 264).

We have been careful to use the term 'statistical' false alarm because even if chance (statistical noise) is an unlikely explanation for the alarm, this does not necessarily mean that there is a problem with quality of care. Data problems and residual confounding are possible reasons for a signal and could be the cause of a 'non-statistical' false alarm.

Local investigations of signals subsequently attributed to non-statistical false alarms are not necessarily a waste of time. For example, signals subsequently attributed to data problems provide opportunities for learning to improve the quality of data. There is some evidence that identification of data problems and feedback can improve the quality of data provided to large databases [22]. Similarly, signals subsequently attributed to residual confounding provide opportunities to learn about patterns of referral of particular types

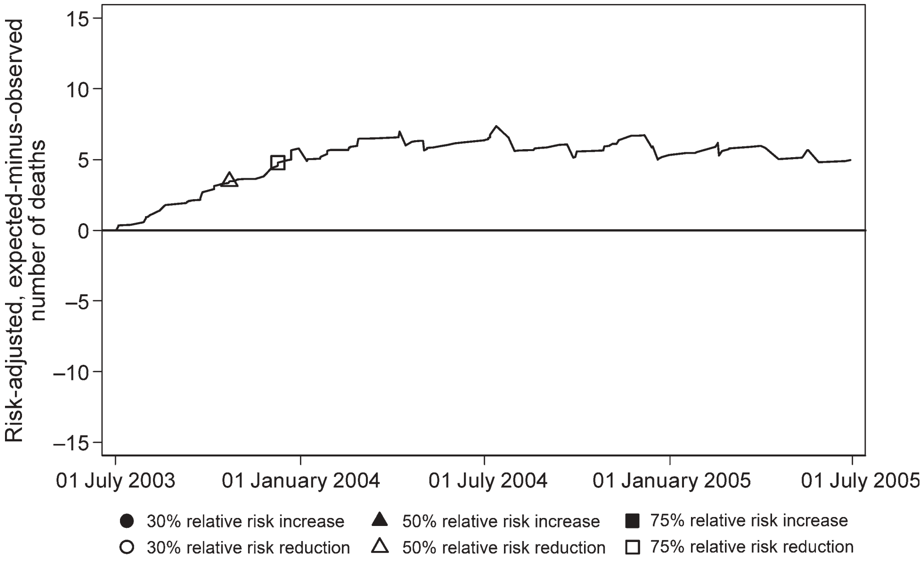

Figure 4 Risk-adjusted, expected-minus-observed plot, Hospital C.

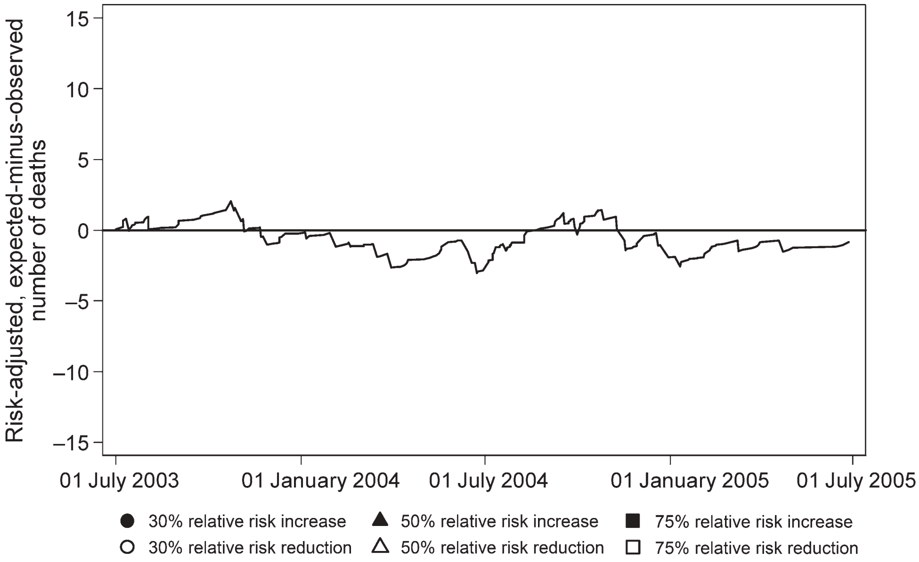

Figure 5 Risk-adjusted, expected-minus-observed plot, Hospital D.

of patients to different hospitals. This is particularly important in a jurisdiction like the state of Queensland, which has small, widely dispersed communities outside the major cities.

Most published reports of control charting in health-care have been based on relatively small clinical data sets [12, 13, 23–25]. There have been a couple of examples of control charting based on larger data sets. For example, Aylin *et al.* demonstrated their use in a pilot study for monitoring mortality rates in primary health-care [26] and Speigelhalter *et al.* conducted a retrospective analysis of three longitudinal datasets to demonstrate the use of risk-adjusted sequential probability ratio tests, a type of log-likelihood-ratio CUSUM [27]. Nevertheless, routine reporting, based on administrative data, is still almost invariably based on cross-sectional analyses [28].

Limitations of the comparison

The signals from the two methods that were compared in this study were not calibrated in the same way. The control charts were based on average run lengths that depend, among other things, on likelihood ratios. Using likelihood has advantages for sequential monitoring and has been used for control charts in industrial settings since the 1950s [29]. One problem with likelihood methods, in general, is that the strength of statistical evidence cannot be easily translated into a probability [30]. This is not an important problem for continuous sequential monitoring, where it is better to use average run lengths to true or false alarm, rather than probability (Type 1 and Type 2 error rates) to describe the characteristics of the plot.

Although, strictly speaking, we did not compare like with like, we used the method of statistical inference that is most commonly used for each type of chart: likelihood for the control charts and frequentist confidence intervals for the cross-sectional plots.

Another limitation of the comparison is that our study did not address the question of which one of the two charts is preferred by hospitals. More work is needed to how best to engage end-users of clinical indicators [4].

Limitations of control charts

One difficulty with control charts is selecting the value of *h* where the chart is said to signal. For clinical trials, the convention, admittedly arbitrary, is to set the Type 1 error rate to 0.05 (or less commonly 0.01) and the Type 2 error rate to 0.20 or 0.10, corresponding to power of 80 or 90%. Unfortunately, there are no similar conventions to guide the selection of values for the average run length to true or false alarm, which might inform the selection of values for *h*. In this paper, we specified the average run length to true alarm to be 100 and found that 50% of hospitals flagged for a relative risk increase of 30 and 28% for a relative risk increase of 75%. We made an explicit decision in Queensland to have more rather than fewer flags because we wanted to be sure of identifying true flags and were tolerant of the costs of investigating false flags.

There is an inherent trade-off between sensitivity and false alarms in any monitoring system [31]. Charts with fewer signals could be obtained by increasing the value of *h* and consequently the average run length to both true and false alarm. Given the incidence and cost of adverse events [32], we prefer a highly sensitive monitoring system. With time, some arbitrary conventions are likely to be developed to guide selection of values for average run length to true and false alarm, just as they have for *P*-values and statistical power in clinical trials.

Another potential problem with control charting is seasonal variation in the expected risk (Y_n) of death. The control charts in this paper were updated monthly, using a risk adjustment model based on the data for that month and the previous 11 months. It might be better to use a risk adjustment model that adjusts for seasonality. This is another example of possible residual confounding and the potential effects might be reduced if the variable 'season' was included in the risk-adjustment model.

The mortality rates for our data on acute myocardial infarction did show variation by month but it did not follow any particular seasonal pattern. For example, for the two financial years studied, the two months with the highest mortality rates were April (15.6%) and August (14.2%); the 2 months with the lowest were July (10.9%) and December (10.3%). More work is needed on how best to update expected values in control charts.

Conclusion

Control charts are not confirmatory statistical tools, but more closely resemble exploratory data analysis [10]. This fits with methods of learning to improve quality called pragmatic science, recently promoted by Berwick [33]. Briefly, this involves tracking effects over time, especially with graphs and then developing and testing theories for improvement using small samples and short experimental cycles. This meshes nicely with a framework where a coordinating agency produces control charts based on administrative data and then refers any unusual sequence of outcomes for more in-depth self-monitoring at a local level.

One strength of control charts is that they can detect problems early. The sooner a potential problem is flagged, the easier it is to correct and to limit the risk to patients and professional reputations [13]. The uncovering of several months (or even years) of poor results, as cross-sectional analyses are designed to do, is potentially distressing for hospitals and patients and can lead to recriminations and blame, rather than learning and improvement. Control charts provide an understandable and up-to-date overview that allows detection of runs of good or bad outcomes and can encourage local investigation and learning.

References

1. Lilford R, Mohammed M, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004;**363**:1147–54.
2. Iezzoni L. Assessing quality using administrative data. *Ann Intern Med* 1997;**127**:666–74.
3. Wright J, Bradley C, Sheldon T, Lilford R. Trial by media: dangers of misinterpretation of medical statistics. *Lancet* 2006;**367**:1139–40.
4. Marshall M, Romano P. Impact of reporting hospital performance. *Qual Saf Health Care* 2005;**14**:77–8.
5. Bolsin S. Routes to quality assurance: risk adjusted outcomes and personal professional monitoring. *Int J Qual Health Care* 2000;**12**:367–9.
6. Shearer A, Cronin C, Feeney D. *The State of the Art of Online Hospital Pubic Reporting: A Review of Forty-Seven Web-sites*. Easton: Delmarva Found, 2004.
7. Agency for Healthcare Quality and Research (AHRQ). *National Quality Measures Clearinghouse*. 2006 http://www.qualitymeasures.ahrq.gov/browse/measureindex.aspx (cited December 2006).
8. National Centre for Classification in Health. *The International Statistical Classification of Diseases and Related Health Problems*, 10th Revision, Australian Modification. Sydney: NCCH, University of Sydney, 2002.
9. Tu J, Austin P, Naylor C, Iron K, Khang H. Acute myocardial infarction outcomes in Ontario. In: Naylor C, Slaughter P (eds). *Cardiovascular Health and Services in Ontario: An ICES Atlas*. Toronto: Institute for Clinical Evaluative Sciences, 1999:83–100.
10. Woodall W. The use of control charts in health-care and public-health surveillance. *J Qual Technol* 2006;**38**:89–104.
11. Grigg O, Farewell V, Spiegelhalter D. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Meth Med Res* 2003;**12**:147–70.
12. Poloniecki J, Valencia O, Littlejohns P. Cumulative risk adjusted mortality chart for detetcing changes in death rate: observational study of heart surgery. *BMJ* 1998;**316**:1697–700.
13. Lovegrove J, Valencia O, Treasure T, Sherlaw-Johnson C, Gallivan S. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997;**350**:1128–30.
14. Spiegelhalter D. Monitoring clinical performance: a commentary. *J Thorac Cardiovasc Surg* 2004;**128**:820–2.
15. 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.
16. Alter D, Naylor C, Austin P, Tu J. Effects of socioeconomic status on access to invasive cardiac procedures and on mortality after acute myocardial infarction. *New Engl J Med* 1999;**341**:1359–67.
17. Rogers C, Reeves B, Caputo M, Ganesh J, Bonser R, Angelini G. Control chart methods for monitoring cardiac surgical performance and their interpretation. *J Thorac Cardiovas Surg* 2004;**128**:811–9.
18. Steiner S, Cook R, Farewell V, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000;**1**:441–52.
19. Deming W. *Out of the Crisis*. Massachusetts: Massachusetts Institute of Technology, 1986.
20. Marshall E, Spiegelhalter D. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998;**317**:1701–4.
21. Spiegelhalter D. Funnel plots for comparing institutional performance. *Stat Med* 2005;**24**:1185–202.
22. Fine L, Keogh B, Cretin S, Orlando M, Gould M. How to evaluate and improve the quality and credibility of an outcomes database: validation and feedback study on the UK cardiac surgery experience. *BMJ* 2003;**326**:25–8.

23. Morton A, Whitby M, McLaws M. The application of statistical process control charts to the detection and monitoring of hospital acquired infections. *J Qual Clin Pract* 2001;**21**:112–7.

24. de Leval M, Francois K, Bull C. Analysis of a cluster of surgical switch failures: application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994;**107**:914–23.

25. Bolsin S, Colson M. The use of CUSUM in the assessment of competence in procedural health care. *Int J Qual Health Care* 2000;**12**:433–8.

26. Aylin P, Best N, Bottle A, Marshall C. Following Shipman: a pilot system for monitoring mortality rates in primary care. *Lancet* 2003;**632**:485–91.

27. Spiegelhalter D, Grigg O, Kinsman R, Treasure T. Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery. *Int J Qual Health Care* 2003;**15**:7–13.

28. Robinowitz D, Dudley R. Public reporting of provider performance: can its impact be made greater? *Ann Rev Public Health* 2006;**27**:517–36.

29. Page E. Continuous inspection schemes. *Biometrika* 1954;**41**:100–15.

30. Blume J. Likelihood methods for measuring statistical evidence. *Stat Med* 2002;**2002**:2563–99.

31. Lim T. Statistical process control tools for monitoring clinical performance. *Int J Qual Health Care* 2003;**15**:3–4.

32. Ehsani J, Jackson T, Duckett S. The incidence and cost of adverse events in Victorian hospitals. *Med J Aust* 2006;**184**:551–5.

33. Berwick D. Broadening the view of evidence-based medicine. *Qual Saf Health Care* 2005;**14**:315–6.

Accepted for publication 1 November 2007