Control chart methods for monitoring cardiac surgical performance and their interpretation

Chris A. Rogers, PhD^{a,b}
 Barnaby C. Reeves, DPhil^c
 Massimo Caputo, MD^a
 J. Saravana Ganesh, FRCS^b
 Robert S. Bonser, FRCS^b
 Gianni D. Angelini, FRCS^a

See related articles on pages 807, 820, 823, and 907.

For more than a decade, there has been increasing interest in monitoring the quality of cardiac surgical performance, as demonstrated by public dissemination of surgeon-specific mortality for coronary artery bypass grafting (CABG) in *The New York Times*,¹ introduction of clinical governance strategies into the United Kingdom National Health Service,² mounting pressure for open scrutiny of results after publication of the Bristol Royal Infirmary Inquiry Panel report,^{3,4} and numerous applications of quality control methods in medicine, both to monitor individuals' results⁵⁻¹⁵ and to compare the performance of individuals or institutions.¹⁶⁻¹⁸ Quality is seen as important not only because of its potential to detect unacceptable surgical results, but also because of the need to ensure quality when training the next generation of surgeons in a high-risk specialty.

All processes, including all aspects of medical care, are assumed to be subject to intrinsic random (*common-cause*) variation. The purpose of quality control charts is to distinguish between random variation and *special-cause* variation, which arises from factors extrinsic to the process. Reducing random variation for a process that is in control requires changing the process itself. Reducing special-cause variation requires identifying factors that cause the process to go out of control and taking appropriate corrective action.

A quality control chart can take one of several forms, depending on the type of data (continuous, binary, or count data—eg, blood loss or length of hospital stay [continuous data], mortality [binary data], or complications [count data]), the quantity of interest (eg, average performance or variability in performance), and the primary objective of the monitoring procedure. Shewhart control charts, for example, were designed for monitoring batches of results.¹⁹ In the surgical context, a *batch* might be a series of operations performed over a period of time. Although these charts have been applied in cardiac surgery,^{5,16} their value for ongoing monitoring of individual results is limited, particularly for low-volume procedures.

Another type of control chart is the cumulative sum (CUSUM). It can be updated after each procedure, is applicable to outcomes for individual surgeons, and provides a method of real-time monitoring of performance. CUSUM charts are based on sequential monitoring of cumulative performance over time and are the focus of this article. Initially developed by Page²⁰ in an industrial context, they have been shown to be most suited for detecting small, persistent process changes.²¹ Williams and colleagues²² first proposed their use in a medical context, and de Leval and associates⁶ were the first to illustrate their ability to detect a cluster of deaths after the arterial switch repair for transposition of the great arteries. Although CUSUM charts are simple to construct, care is needed to avoid overinterpreting or misinterpreting them.

The purposes of this article are to (1) describe different forms of CUSUM charts for monitoring performance over time when the outcome of interest is binary (eg, mortality or cardiac-related events),^{6,22} (2) explain how the charts should be interpreted, (3) highlight frequent misunderstandings, and (4) recommend ways the

From the Bristol Heart Institute, University of Bristol, Bristol, United Kingdom,^a the UK Cardiothoracic Transplant Audit Steering Group, Clinical Effectiveness Unit, The Royal College of Surgeons of England, London, United Kingdom,^b and Health Services Research Unit, London School of Hygiene & Tropical Medicine, London, United Kingdom.^c

This study was performed with funding from the British Heart Foundation, the Garfield Weston Trust, and the United Kingdom Department of Health.

Received for publication Sept 19, 2003; revisions requested March 4, 2004; revisions received March 10, 2004; accepted for publication March 16, 2004.

Address for reprints: G. D. Angelini, FRCS, Bristol Heart Institute, University of Bristol, Bristol Royal Infirmary, Bristol BS2 8HW, United Kingdom (E-mail: g.d.angelini@bristol.ac.uk).

J Thorac Cardiovasc Surg 2004;128:811-9
 0022-5223/\$30.00

Copyright © 2003 by The American Association for Thoracic Surgery

doi:10.1016/j.jtcvs.2004.03.011

charts should be used. We also consider extensions of the CUSUM chart that control for case mix: variable life-adjusted displays (VLAD,⁷ also called cumulative risk-adjusted mortality [CRAM] plots⁸), and the risk-adjusted sequential probability ratio test (SPRT).⁹ We describe the parameters needed to construct the charts, their control limits, and alternative graphical presentations of data. We focus on binary outcomes because they are used to monitor cardiac surgery performance. The methods are illustrated by using two example data sets: a single United Kingdom hospital database of cardiac operations and a national database of cardiothoracic transplantations in the United Kingdom.

Data Sets

Cardiac Surgery Database

The Bristol Heart Institute has prospectively collected a standard set of data on all adult cardiac procedures since April 1996.^{23,24} Data used for illustration comprise 1372 elective and urgent CABG procedures performed between April 1996 and September 2002.¹⁰ All operations were performed by the lead academic consultant or one of four residents. The outcome chosen for performance monitoring was surgical failure, defined as the occurrence of one or more of 11 cardiac-related events.¹⁰ Overall failures were 8.5% (95% confidence interval, 7%-10%). Multiple logistic regression, applied to the complete data set, was used to identify predictors of failure. The predicted risk of surgical failure for each of the 1372 patients was then estimated from the resulting model. Results are presented for a subset of off-pump CABG.

Cardiothoracic Transplantation Database

A national clinical database of cardiothoracic transplantations and outcomes was established in April 1995, and all 8 centers in the United Kingdom that perform these procedures have contributed data since then. Data returns are in excess of 95%, and all data are subject to rigorous validation.²⁵ The data used for illustration comprise 1341 adult orthotopic heart transplantations performed between July 1995 and September 2002. The outcome chosen for monitoring was 30-day postoperative mortality, which was 12% (95% confidence interval, 10%-14%). Multiple logistic regression analysis was used to identify predictors of mortality for the July 1995 to March 2001 cohort ($n = 1173$), and the model was evaluated using subsequent transplantations (April 2001 to September 2002; $n = 168$). Details of the risk factors considered and model development are available on request.

Constructing and Interpreting Charts

Non-Risk-Adjusted Methods

Choice of outcome, the event against which performance is being measured, varies depending on context. Throughout we shall simply refer to an unsuccessful outcome as a *failure* and a successful outcome as a *success*. We shall also focus mainly on detecting an increase in failures, although the methods are equally applicable for detecting their reduction. Although our emphasis will be on risk-adjusted control charts, before introducing them, we will discuss and illustrate non-risk-adjusted charts.

Cumulative failure charts. The simplest and most intuitive form of CUSUM chart is a graph of the cumulative (total) number of failures (on the vertical axis) against operation number (on the horizontal axis; *stepped lines* in Figure 1). As each operation is performed and outcome assessed, the cumulative number of failures either remains unchanged if a success occurs (and the graph continues horizontally) or is incremented by 1 if a failure occurs (and the graph rises). The graph has an immediate visual interpretation, because an increase in gradient (slope) indicates more frequent failures. However, it is of limited value without control boundaries to indicate whether an increase in gradient is consistent with a process going out of control (ie, a genuine increase in the failure rate) or with simple random variation. The control boundaries illustrated are derived from the SPRT²⁶ and are constructed to test the null hypothesis (H_0) that the failure rate is p_0 , against the alternative (H_1) that the failure rate has increased to p_1 .

To construct control boundaries, 4 parameters must be specified: (1) risk of failure when the process is in control (acceptable failure rate; p_0); (2) failure rate considered unacceptable (p_1 , where $p_1 > p_0$); (3) α , the probability of concluding that the failure rate has increased when, in fact, it has not (false-positive, or type I error); and (4) β , the probability of concluding that the failure rate has not increased when, in fact, it has (false-negative, or type II error). Choices of α and β depend on the application and relative costs of false-positive and false-negative conclusions; they are commonly set to .10 (10%), .05 (5%), or .01 (1%).⁹ Given values for p_0 , p_1 , α , and β , the upper and lower control limits (or boundary lines)— l_1 and l_0 , respectively—are constructed according to formulas given in Appendix 1 (*dashed lines*, Figure 1).

It is a common misconception that a CUSUM graph that remains within control boundaries constitutes evidence that the process is in control. If the graph of cumulative failures crosses the upper boundary, l_1 , then we conclude that the failure rate has increased to the unacceptable rate, p_1 . If it crosses the lower boundary, l_0 , we conclude that the failure rate is less than or equal to the acceptable rate, p_0 . When a graph remains between these boundaries, the evidence remains inconclusive, and monitoring should continue (Figure 1). The natural progression of the graph for an individual or institution with acceptable performance is toward the lower boundary for this method of constructing control limits.

Upper and lower boundary lines are always parallel. Their slope s (Appendix 1) is not directly interpretable. It depends on the values of p_0 and p_1 ; the closer p_1 is to p_0 (ie, the smaller the increase to be detected), the smaller s is and the shallower the slope. The points at which the lower and upper boundary lines intersect with the vertical CUSUM axis (h_0 and h_1 ; Appendix 1) are determined by α and β , which are typically set to the same value. The smaller the values for α and β , the higher the upper boundary and the lower the lower boundary. It is common for 2 sets of boundary lines to be included on the chart, corresponding to different choices for α and β . Lines for the higher values of α and β are often referred to as *alert* lines, with lower values of α and β defining the *alarm* or *action* lines. The distance between the boundary lines ($h_0 + h_1$) also depends on the odds ratio (or, equivalently, p_0 and p_1 ; Appendix 1); the smaller the odds ratio, the greater the distance between the boundaries (for a given choice

Figure 1. Cumulative failure charts for (a) surgical failure after off-pump CABG (OPCAB) and (b) 30-day mortality after orthotopic heart transplantation in adults. Expected failure rates (p_0) were set at overall failure rates for the programs as a whole: (a) 8.5% for 1 consultant and 4 residents and (b) 12% for 8 centers. Boundary lines were constructed to detect a 50% increase in failures (odds ratio, 1.5): (a) 3.7% ($p_1 = 12.2\%$) and (b) 5.0% ($p_1 = 17.0\%$). False-positive (α) and false-negative (β) error rates are 5% for both charts. Lines representing expected cumulative failures (— · —) are shown in both charts, although these are not usually included. In (a), which depicts the consultant and 1 of the 4 residents, the consultant's failure rate is similar to the overall failure rate (closely follows the — · — line), but is less than expected for the resident. The resident's performance was confirmed as acceptable (or better) after 100 operations, when the lower boundary line was reached. In (b), which depicts 2 of the 8 transplant centers, performance at center A was consistently better than expected and was confirmed to be acceptable or better after 80 transplantations. Performance at center B was in line with overall mortality for the first 100 transplantations, but increased steadily thereafter. By transplantation 167, center B was close to the 5% upper boundary, having already crossed the 10% upper boundary (not shown).

of α and β) and the longer the sequence of operations needed before a conclusion is reached.

Cumulative log-likelihood ratio chart. An alternative but equivalent presentation of the data involves graphing a modified CUSUM against the operation number (Figure 2). As with the cumulative failures graph, the sum starts at 0, but is then incremented by $1 - s$ for a failure and decremented by s for a success. The value of s is defined by p_0 and p_1 (Appendix 1). Boundary lines are horizontal, and their position on the chart is defined by h_0 and h_1 . Interpretation of the graph in relation to the boundary lines is the same as for the cumulative failures chart. If performance is acceptable, the graph will tend downward toward the lower boundary; it will not follow the horizontal axis.

Cumulative observed minus expected failure graph. It is possible to construct a chart so that acceptable performance gives rise to a graph that oscillates around a horizontal axis (Figure 3). This type of chart requires the expected value for the CUSUM to be 0 if the process is in control. The graph starts at 0, but is incremented by $1 - p_0$ for a failure and decremented by p_0 for a success. This

graph is more intuitive because it is easier to identify changes in the failure rate: the graph moves upward if the failure rate increases and downward if it decreases. To test the hypothesis that the failure rate has increased from p_0 to p_1 , boundary lines would need to be drawn sloping upward with the gradient $s - p_0$. Drawing horizontal boundary lines (h_0 and h_1) would represent a change of hypothesis being tested, ie, a change of p_0 and p_1 , and the horizontal axis would no longer represent acceptable performance. If boundary lines are drawn on a cumulative observed minus expected failure chart, care needs to be taken to specify clearly the hypothesis being tested and to calculate appropriate boundary lines.

Choice of charts. These different formats of charts are equally valid; choice is largely a matter of personal preference. The chosen format needs to be specified, and if boundary lines are included, they must be accompanied by an explanation of their construction and the underlying hypothesis being tested.

Cumulative observed minus expected failure graphs are intuitive because changes in gradient are more immediately apparent.

Figure 2. Cumulative log-likelihood ratio test charts for (a) surgical failure after off-pump CABG (OPCAB) and (b) 30-day mortality after orthotopic heart transplantation in adults. Data and parameter settings for constructing boundary lines (p_0 , p_1 , α , and β) are the same as for Figure 1. Lines representing expected cumulative failures have not been included; note that such lines, if included, would not be horizontal through 0 but would slope downward from 0 toward the lower boundary, which denotes acceptance of H_0 . These figures provide an alternative representation of the data shown in Figure 1. Interpretation of the graphs in relation to the boundary lines is the same. The points at which the graphs for the resident and center A cross the lower acceptance boundary coincide with Figure 1.

However, plotting boundary lines to detect deviations from acceptable performance is more intuitive with cumulative failures or cumulative log-likelihood ratio charts. Therefore, we consider the two types of chart to be complementary.

A line with a gradient corresponding to the acceptable (expected) failure rate could be added to cumulative failure charts, but it would not run parallel to the boundary lines. It is important to distinguish between interpretation of the graph relative to the boundary lines and interpretation in relation to the acceptable failure rate (Figure 1).

Risk-Adjusted Methods

The methods just described have been extended to adjust or control for case mix in sequential monitoring of health outcomes. The concept is simple. Rather than assuming that the acceptable failure rate is the same for all patients, the predicted risk of failure is allowed to vary among individuals. Accepted statistical models (eg, those based on the Parsonnet score^{8,9} or EuroSCORE²⁷) or empirically derived models^{6,10,17} are used to estimate the patient-specific predicted probability of failure. The risk-adjusted SPRT chart⁹ is the risk-adjusted analog to the cumulative log-likelihood ratio chart. VLAD⁷ or CRAM charts^{8,27} are constructed on this principle and are analogous to the cumulative observed minus expected failure chart. Advantages and disadvantages of different forms of unadjusted charts apply equally to their risk-adjusted counterparts.

VLAD or CRAM chart. The graph, which starts at 0, is incremented by $1 - p_{0i}$ for a failure and is decremented by p_{0i} for a success, where p_{0i} denotes the predicted probability of failure for operation i , derived from the appropriate risk model (Figure 4). The graph has a natural interpretation: it moves upward if the failure rate increases above that predicted by the risk model, moves downward if the rate decreases, and oscillates around 0 if performance is consistent with predicted risks, ie, acceptable. Although changes in gradient are easy to see, constructing boundary lines is not straightforward. Methods for detecting changes have been proposed,^{7,8} but they do not equate to a hypothesis test in quite the same way as described for CUSUM charts.

Risk-adjusted SPRT chart. The risk-adjusted analog of the CUSUM chart, with boundary lines based on a SPRT, was first described in a medical context by Spiegelhalter and colleagues.⁹ The risk-adjusted cumulative log-likelihood ratio statistic is used, and boundary lines are drawn horizontally (Figure 4). The graph starts at 0 and is incremented by $1 - s_i$ for a failure and decremented by s_i for a success. The value of s_i is defined by the predicted risk of failure for operation i (p_{0i}) and the increase in risk that the chart is designed to detect.

For the unadjusted chart, increase in risk is defined in terms of the unacceptable failure rate. However, when risk for each patient varies, it does not make sense to have a common unacceptable rate applied across all operations; it needs to vary according to the predicted risk of failure for the procedure. This variable unaccept-

Figure 3. Cumulative observed minus expected failure charts for (a) surgical failure after off-pump CABG (OPCAB) and (b) 30-day mortality after orthotopic heart transplantation in adults. Data are the same as in Figures 1 and 2. Expected failure rates were set at the overall failure rates for the programs as a whole: (a) 8.5% and (b) 12.0% (Figure 1). This figure represents another alternative to Figure 1. The horizontal (— · · —) line represents the expected failure rate. The graph moves upward if there are more failures than expected (eg, center B) and downward if there are fewer (eg, resident surgeon and center A). Overall, the consultant's failure rate was as expected, and the resident had fewer failures than expected. Center A also had many fewer deaths than anticipated, but center B reported an excess.

able rate is achieved by defining the increase in terms of a relative risk (ie, odds ratio), rather than a specific rate. An odds ratio of 2, for example, would equate approximately to a doubling of patient-specific risk of failure, an odds ratio of 1.5 to a 50% increase in failure risk, and so on. The natural progression of the risk-adjusted graph for an individual or institution with acceptable performance is toward the lower boundary.

Choice of charts. The VLAD or CRAM chart and risk-adjusted SPRT chart are complementary and designed to account for case mix. The VLAD chart is intuitive, because the horizontal axis corresponds to expected outcome, and if performance is in line with expectations, the chart should oscillate around 0. A change in gradient, indicating a process that may be going out of control, is easily spotted. Its disadvantage is that boundary lines are not easily constructed.

In contrast, the SPRT chart has no intuitive interpretation, but it has the advantage of providing a formal test of an explicit hypothesis. Although either chart is preferred to the unadjusted CUSUM for applications in which case-mix adjustment is appropriate, their usefulness is only as good as the ability of the risk model to accurately predict the outcome for different patient profiles. No risk model is perfect; none can completely adjust for all factors that influence outcome.

Crossing Boundary Lines

If the graph (either risk adjusted or not) crosses the upper boundary line, then H_0 is rejected, and performance is confirmed to have reached the predefined unacceptable level. In this situation, the individual or team should investigate the cause of the unacceptable performance, implement changes as necessary, and resume monitoring. If performance improves thereafter, the graph will start to decline and return to the “continue monitoring” zone. When the converse occurs and the graph crosses the lower acceptance boundary, we suggest that it be reset to 0 before monitoring is resumed, thereby increasing the sensitivity of the monitoring procedure by avoiding buildup of excessive “credit.”⁹

Discussion

CUSUM charts in their various forms are simple to construct and easy to interpret when key parameters are defined correctly. The two main forms—the cumulative sum and cumulative observed minus expected failure graph—are complementary. Risk-adjusted versions are available and should be used when case-mix adjustment is appropriate, that is, when the population is heterogeneous and diverse outcomes may be anticipated. A robust, validated, highly

Figure 4. Risk-adjusted cumulative observed minus expected failure (VLAD/CRAM) and cumulative log-likelihood ratio test (SPRT) charts for (a) surgical failure after off-pump CABG (OPCAB) and (b) 30-day mortality after orthotopic heart transplantation in adults. Data are the same as for Figures 1 to 3. Patient-specific expected failure rates (p_{0i}) were estimated from empirically derived risk models. Boundary lines were constructed to detect an increase in failure rate equivalent to a 50% increase in risk (odds ratio, 1.5) in both cases. False-positive (α) and false-negative (β) error rates are 5% for both charts. The graphs are interpreted exactly as the unadjusted graphs. Risk adjustment has had little effect for the OPCAB data. From operation 100 forward, the consultant's results matched expectations almost exactly, and the resident's risk-adjusted SPRT falls just short of the lower boundary. In contrast, risk adjustment reduced the apparent excess deaths at center B by approximately half, and the risk-adjusted SPRT remains firmly in the "continue monitoring" zone. For center A, the effect of risk adjustment is minimal, and performance is clearly better than expected. If monitoring had been restarted each time the graph crossed the lower acceptance boundary (see "Discussion"), the graph for center A would have crossed the lower boundary for a second time after the 178th transplantation and for a third time after the 269th.

discriminating model of risk should be used; a model that is poorly calibrated or has poor discrimination will provide inadequate adjustment for case mix. However, no case-mix adjustment will remove all confounding effects. An alternative approach to case-mix adjustment is to restrict the analysis to a relatively homogeneous group of patients,⁵ but this could result in poor performance going undetected if a surgeon's results for the monitored subgroup are in line with what is expected, but are suboptimal for subgroups excluded from analysis.

When adding alert or alarm lines to a chart, the user needs to be clear about the role of the different parameters used in constructing the lines (ie, the hypothesis under test),

and values of these parameters must be specified clearly for the reader. There are literature examples that are misleading in this respect. Williams and colleagues,²² for example, suggest that the parameter s , termed the *reference* or *target* value, should be chosen to reflect the expected or acceptable failure rate, that p_1 defines the unacceptable rate, and that p_0 is chosen "by trial and error" to give the required s . Although these choices provide a directly interpretable cumulative observed minus expected failure chart, the boundary lines correspond to a test of p_0 (the "trial and error" value) versus p_1 , not of s versus p_1 . This misunderstanding of the role and interpretation of p_0 and s has been perpetuated in the literature. Novick and colleagues,^{11,12} in assessing out-

comes after on- and off-pump CABG, use a similar approach, but alert and alarm lines are drawn at an angle rather than horizontally. They could be using a hybrid cumulative observed minus expected failure chart with appropriate boundary lines, but the text would suggest not, because they quote the formulas for a cumulative failures chart. Furthermore, in constructing the boundary lines, they fail to state the hypothesis being tested, specify the value for p_0 , and apparently set p_1 to the maximum acceptable failure rate (10%). Any inferences made from the graphs in relation to boundary lines are spurious and misleading. In plotting the CUSUM, they also restricted the graph to nonnegative values.¹² This leads us to further question the chart's validity, because cumulative observed minus expected failure graphs have both positive (more failures than expected) and negative (fewer failures than expected) values. The rationale for restricting the graph in this way is not explained.

The choice of values for α and β , the false-positive and false-negative rates, as well as those for p_0 and p_1 (or, equivalently, the odds ratio), are not always immediately obvious and depend on the setting and specific application. Spiegelhalter and colleagues⁹ suggest that if the charts are used to monitor several surgeons or institutions, it may be appropriate to use smaller α and β , thereby making the boundaries more stringent and reducing the probability that the graph for a surgeon or institution performing normally crosses the boundary by chance. They suggest a Bonferroni-type correction, replacing α and β by α/n and β/n , where n is the number of surgeons or institutions.

When a graph crosses a boundary, a conclusion regarding performance (as expected or otherwise) is reached. If, as suggested by Spiegelhalter and colleagues,⁹ the graph is reset to 0 after it crosses the lower acceptance boundary, to avoid building up excessive "credit" and increasing the sensitivity of the monitoring procedure, strict interpretation of α and β is lost. Successive restarts increase the overall chance of a false-positive result and decrease the chance of a false-negative result.⁹

The concept of the *average run length* is a feature of sequential monitoring schemes that use boundary lines to assess performance. Average run length is the average number of observations or operations needed before a conclusion (acceptable vs unacceptable performance) can be drawn (1) when the process is in control and (2) when the process has moved out of control. Average run lengths for the unadjusted CUSUM can be determined by simulation for different choices of chart parameters, but for the risk-adjusted analog, the calculations are less straightforward; run length will depend not only on α , β , and the chosen odds ratio, but also on the risk profile of the study population. If the risk profile changes over time, results based on prior data may not reflect average run length accurately when the chart is used to monitor subsequent performance.

Steiner and colleagues^{28,29} used the concept of average run length to construct boundaries for their risk-adjusted CUSUM charts. Their charts closely resemble the risk-adjusted SPRT chart described here insofar as they use the same form of log-likelihood ratio statistic, but there are key differences: they do not allow the CUSUM to go below 0, and the graph is reset not when the lower acceptance boundary is crossed, but *each time* the sum decreases to less than 0. There is no concept of accepting H_0 and no simple formula for constructing the boundary. The chart's control limit is estimated pragmatically and in a computationally intensive fashion.

There are a number of other monitoring techniques that we have not considered here, such as the sets³⁰ and cumulative score³¹ methods, which have been used for public health monitoring of rare events, and Shewhart^{16,19} and funnel charts.³²

Much research remains to be performed in developing and evaluating sequential monitoring procedures in health care, where case-mix adjustment, however crude, is essential for realistic evaluation and comparison. If the procedure is relatively uncommon and activity is low, a monitoring scheme could run for several years before a conclusion is drawn, which might lead one to question the value of sequential monitoring in such a setting and the relevance of past data to what is happening now. One approach might be to modify the methods so that less importance is given to historic data. Moving averages based on the last n procedures¹³ or that give less weight to increasingly historic cases, such as exponentially weighted moving average (EWMA) charts, are approaches to accomplish this.⁶ However, as far as we are aware, this method has yet to be applied in a risk-adjusted setting, and it is not clear how one would construct appropriate boundary lines. Alternatively, one might increase the number of surgical cases considered by simultaneously monitoring several different procedures performed by the same surgeon or institution (eg, monitoring the overall cardiothoracic transplantation program, rather than assessing heart and lung programs separately). The appropriateness of doing so would depend on the context, but even if it is sensible from a clinical perspective, it is not clear how best to combine the data computationally.

We have described methods for sequential monitoring that focus on detecting an increase in the failure rate. The lower boundary provides confirmation that the failure rate is acceptable (equivalent to accepting H_0). The methods described apply equally well to the problem of detecting a decrease in failure rate. The charts are constructed in exactly the same way; the only difference is that p_1 (the alternative failure rate) is set to a value less than p_0 (or, equivalently, the odds ratio is set to <1). Graphs constructed in this way

could provide a useful method of confirming or validating the performance of trainees.¹⁰

Recommendations and Limitations

We believe that sequential monitoring tools have a role in assessing and evaluating outcomes, but we suggest that they complement other methods rather than replace them. Although there is still work to be done in evaluating the performance of different monitoring methods, this should not preclude their use. Sequential monitoring charts are simple to construct, and adjustment for case mix is accommodated in a straightforward way. We recommend the following:

- Complementary use of cumulative observed minus expected failure and CUSUM charts.
- Clear specification of the way in which the CUSUM chart is constructed (cumulative failures versus cumulative log-likelihood ratio statistic).
- Careful consideration of choice of values for the parameters used to define boundaries when these are included and explicit statement of these choices to the reader.
- Case-mix adjustment, if available.
- Interpretation of the graphs for the reader (in line with this article), particularly with respect to performance that remains within the boundary lines.

There are, however, a number of caveats:

- Users need to guard against overinterpreting the graphs, particularly the more intuitive cumulative observed minus expected failure graphs without boundary lines.
- Monitoring a health-care process is not the same as monitoring a manufacturing process; case mix represents a fundamental difference, and risk adjustment is imperfect and cannot remove all confounding.
- If monitoring schemes are to be accepted by those whose outcomes are being assessed, an atmosphere of constructive evaluation, not “blaming” or “naming and shaming,” is essential; apparent poor performance could arise for a number of reasons that should be explored systematically.

References

1. Green J, Wintfeld N. Report cards on cardiac surgeons. *N Engl J Med*. 1995;332:1229-32.
2. Scally G, Donaldson L. The NHS's 50 anniversary. Clinical governance and the drive for quality improvement in the new NHS in England. *BMJ*. 1998;317:61-5.
3. Learning from Bristol: the report of the public inquiry into children's heart surgery at the Bristol Royal Infirmary 1984-1995. In: BRI Inquiry Panel. London: The Stationary Office; 2001.
4. Spiegelhalter D, Aylin P, Best N, Evans S, Murray G. Commissioned analysis of surgical performance by using routine data. *J R Stat Soc A*. 2002;165:1-31.
5. Shahian D, Williamson W, Svensson L, Restuccia J, D'Agostino R. Applications of statistical quality control to cardiac surgery. *Ann Thorac Surg*. 1996;62:1351-9.
6. de Leval M, Francois K, Bull C, Brawn W, Spiegelhalter D. Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg*. 1994;107:914-24.
7. 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.
8. Poloniecki J, Valencia O, Littlejohns P. Cumulative risk-adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *BMJ*. 1998;316:1697-700.
9. 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.
10. Caputo M, Reeves B, Rogers C, Ascione R, Angelini G. Monitoring the performance of residents during training in off-pump coronary surgery. *J Thorac Cardiovasc Surg*. 2004;128:907-15.
11. Novick R, Fox S, Stitt L, Swinamer S, Lehnhardt K, Rayman R, et al. Cumulative sum failure analysis of a policy change from on-pump to off-pump coronary artery bypass grafting. *Ann Thorac Surg*. 2001;72:S1016-21.
12. Novick R, Fox S, Stitt L, Kiai R, Swinamer S, Rayman R, et al. Assessing the learning curve in off-pump coronary artery surgery via CUSUM failure analysis. *Ann Thorac Surg*. 2002;73:S358-62.
13. Brown S, Benneyan J, Theobald D, Sands K, Hahn M, Potter-Bynoe G, et al. Binary cumulative sums and moving averages in nosocomial infection cluster detection. *Emerg Infect Dis*. 2002;8:1426-32.
14. Lawrance R, Dorsch M, Sapsford R, Mackintosh A, Greenwood D, Jackson B, et al. Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice: observational study. *BMJ*. 2001;323:324-7.
15. Bolsin S, Colson M. The use of the Cusum technique in the assessment of trainee competence in new procedures. *Int J Qual Health Care*. 2000;12:433-8.
16. Mohammed M, Cheng K, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet*. 2001;357:463-7.
17. Tekkis P, McCulloch P, Steger A, Benjamin I, Poloniecki J. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *BMJ*. 2003;326:786-90.
18. Marshall E, Spiegelhalter D. League tables of in vitro fertilisation clinics: how confident can we be about the rankings? *BMJ*. 1998;316:1701-4.
19. Shewhart W. Economic control of quality of manufactured product. Princeton (NJ): Van Nostrand Reinhold; 1931.
20. Page E. Continuous inspection schemes. *Biometrika*. 1954;41:100-14.
21. Montgomery D. Introduction to statistical quality control. 2nd ed. New York: John Wiley; 1991.
22. Williams S, Perry B, Schlup M. Quality control: an application of the CUSUM. *BMJ*. 1992;304:1359-61.
23. Caputo M, Bryan A, Capoun R, Mahesh B, Ciulli F, Hutter J, et al. The evolution of off-pump coronary surgery in a single institution. *Ann Thorac Surg*. 2002;74:S1403-7.
24. Caputo M, Chamberlain M, Ozalp F, Underwood M, Ciulli F, Angelini G. Off-pump coronary operations can be safely taught to cardiothoracic trainees. *Ann Thorac Surg*. 2001;71:1215-9.
25. Anyanwu A, Rogers C, Murday A. Intrathoracic organ transplantation in the United Kingdom 1995-1999: results from the UK cardiothoracic transplant audit. *Heart*. 2002;87:449-54.
26. Wald A. Sequential tests in industrial statistics. *Ann Math Stat*. 1945;6:117-86.
27. Sergeant P, de Worm E, Meyns B, Wouters P. The challenge of departmental quality control in the reengineering towards off-pump coronary artery bypass grafting. *Eur J Cardiothorac Surg*. 2001;20:538-43.
28. Steiner S, Cook R, Farewell V. Risk-adjusted monitoring of binary surgical outcomes. *Med Decis Making*. 2001;21:163-9.
29. Steiner S, Cook R, Farewell V, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics*. 2000;1:441-52.

30. Chen R. A surveillance system for congenital malformations. *J Am Stat Assoc.* 1978;73:323-7.
31. Wolter C. Monitoring intervals between rare events: a cumulative score procedure compared with Rina Chen's sets technique. *Methods Inf Med.* 1987;26:215-9.
32. Spiegelhalter D. Funnel plots for institutional comparison. *Quality Saf Health Care.* 2002;11:390-1.

Appendix 1

Construction of CUSUM charts with control limits (alert or alarm lines) requires specifying 4 parameters: p_0 , the acceptable event rate; p_1 , the unacceptable event rate we want to detect; α , the type I error rate (probability of rejecting the null hypothesis [H_0] that the event rate is p_0 when it is true); and β , the type II error rate (probability of accepting the null hypothesis [H_0] when it is false).

If X_i denotes the outcome for procedure i , with $X_i = 1$ if a failure occurred and $X_i = 0$ if it did not, then for a graph of the cumulative number of events E_i , where

$$E_i = E_{(i-1)} + X_i \text{ with } E_0 = 0, \quad (1)$$

against operation number i , the upper control boundary or limit l_1 (to detect an increase from p_0 to p_1) and the lower control limit l_0 (to accept p_0) are defined by

$$l_1 = i \times s + h_1 \quad (2)$$

and

$$l_0 = i \times s - h_0, \quad (3)$$

where

$$s = \frac{\ln\left(\frac{1-p_0}{1-p_1}\right)}{\ln(\text{OR})} \quad (4)$$

$$\ln(\text{OR}) = \ln\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}\right) = \ln\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) \quad (5)$$

$$h_0 = \frac{\ln\left(\frac{1-\alpha}{\beta}\right)}{\ln(\text{OR})} \quad (6)$$

and

$$h_1 = \frac{\ln\left(\frac{1-\beta}{\alpha}\right)}{\ln(\text{OR})}, \quad (7)$$

where OR is the odds ratio corresponding to an increase in event rate from p_0 to p_1 .

To construct the same chart, but with horizontal control limits, one graphs

$$T_i = T_{(i-1)} + (X_i - s), \text{ with } T_0 = 0 \quad (8)$$

against operation number i , and then the control limits are given by $-h_0$ (Equation 6) and h_1 (Equation 7).

The corresponding risk-adjusted chart (SPRT) is constructed in a similar manner. The control limits are horizontal and defined as above (for Equations 6 and 7). Cumulative sum T_i in Equation 8 is replaced by

$$T_i^* = T_{(i-1)}^* + (X_i - s_i), \text{ with } T_0^* = 0, \quad (9)$$

where

$$s_i = \frac{\ln\left(\frac{1-p_{0i}}{1-p_{1i}}\right)}{\ln(\text{OR})}, \quad (10)$$

p_{0i} represents the procedure-specific risk-adjusted probability of failure, and p_{1i} represents the procedure-specific risk-adjusted probability of failure under the alternative hypothesis, H_1 , corresponding to an increase in odds ratio of size OR.

In practice, it is not necessary to calculate p_{1i} for each procedure i because it can be shown that

$$\ln\left(\frac{1-p_{0i}}{1-p_{1i}}\right) = \ln[(1-p_{0i}) + \text{OR} \times p_{0i}], \quad (11)$$

so

$$s_i = \frac{\ln[(1-p_{0i}) + \text{OR} \times p_{0i}]}{\ln(\text{OR})} \quad (12)$$

T_i^* (Equation 9) corresponds to the log-likelihood ratio given by Spiegelhalter and colleagues⁹ except for the scale adjustment of both the statistic and the control limits by $\ln(\text{OR})$.

Variable life-adjusted displays (VLAD) or cumulative risk-adjusted mortality (CRAM) charts, in contrast, graph

$$V_i = V_{(i-1)} + (X_i - p_{0i}), \text{ with } V_0 = 0 \quad (13)$$

against i , where p_{0i} is defined as described previously. Replacing the predicted p_{0i} by the constant acceptable event rate, p_0 , provides an equivalent unadjusted graph.