Monitoring cardiac surgical performance: A commentary

Tom Treasure, MS, MD, FRCS^a

Stephen Gallivan, PhD^b

Chris Sherlaw-Johnson, MSc, FORS^b

See related article on pages 807, 811, 820, and 907.

Marc de Leval¹ presented a cumulative sum (CUSUM) chart to The American Association for Thoracic Surgery in Chicago 10 years ago. His graph (Figure 1) showed an outstanding series of 52 cases performed with just 1 death early in the adoption of the arterial switch operation for transposition of the great arteries. The CUSUM graph concluded with a similarly excellent, nearly flat line with 1 death in the most recent 39 cases. Sandwiched between them was a cluster of deaths. de Leval's CUSUM chart was simple, explicit, and intuitive; each operation moved the graph 1 unit along the horizontal axis, and each death moved it up by 1 unit on the vertical axis. It enabled those of us fortunate to be present at what proved to be a landmark presentation to follow the story with absolute clarity. de Leval charted the results for a single procedure. Prompted by his presentation, and convinced that this method would help us display and understand our outcomes better, we worked toward a method of displaying data sequentially that would also allow for variable risk in series of different case mix. We dubbed the method *variable life-adjusted display* (VLAD),² but other terms have also been used.³

Rogers and colleagues⁴ discuss use of this and other charts in a timely and helpful tutorial. VLAD charts are now widely used. The Society of Cardiothoracic Surgeons of Great Britain and Ireland (SCTS) displays outcome in this format,⁵ and the charts are commonly used to inspect trends in mortality for surgical units, individual surgeons, and trainees. As Rogers and colleagues⁴ point out, control charts have the merit of being intuitive. Principles of the display are readily grasped. Upward and downward trends and their duration and steepness can be read instantly and clearly. However, these charts are methods of displaying data; they are not tests of statistical significance. Even so, statistical methods have been developed that assist their interpretation. We have suggested a method for use in cardiac surgery (Figure 2) to assess variation in outcomes across a series of cases,⁶ and Grunkemeier and colleagues³ have used 95% two-sided prediction limits. Spiegelhalter included alert and alarm lines in de Leval's original display.¹

There is a notion that poor performance is always known on the inside. If so, by whom? The anesthesiologist or perfusionist? Let us illustrate the fallacy of that belief and the power of sequential displays. Figure 3 (plotted after the event) displays results of several surgeons performing coronary artery bypass grafting. Risk adjustment was by the original Parsonnet score.⁷ The surgeon whose results are graphed in bold was in line with the other surgeons for the first 100 operations, matched those expected on the basis of the Parsonnet score for the next 100, and was obviously worse than expected for the third 100. He was perceived as being at the peak of his intellectual and technical powers and known as a surgeon willing to take on the most exacting cases. The excess deaths went unnoticed by him and his colleagues. Then, he was suddenly discovered to have an unsuspected cortical visual defect.

It should be noted that there are runs of good and not so good results within the other surgeons' series and in any VLAD plot.²⁻⁴ These performance variations are in themselves interesting and appear greater than those seen when we simulate with randomly allocated deaths within a sequence. That is to say, they are more than *common-cause* variation. However, the particular downward trend shown in Figure

From the Department of Thoracic Surgery, Guy's Hospital,^a and the Clinical Operational Research Unit, Department of Mathematics, University College,^b London, United Kingdom.

Received for publication Feb 23, 2004; accepted for publication March 4, 2004.

Address for reprints: Tom Treasure, MS, MD, FRCS, Consultant in Thoracic Surgery, Guy's Hospital, St Thomas's St, London SE1 9RT, United Kingdom (E-mail: Tom.Treasure@ukgateway.net).

J Thorac Cardiovasc Surg 2004;128:823-5
0022-5223/\$30.00

Copyright © 2004 by The American Association for Thoracic Surgery

doi:10.1016/j.jtcvs.2004.03.023

Figure 1. Cumulative failures (CUSUM) according to operative sequence (patient number) among 104 consecutive neonatal arterial switch repairs of transposition of the great arteries. Heavy solid line is CUSUM, fine solid lines are 80% alert lines, and dashed lines are 95% alarm lines. Cluster of deaths is apparent in midportion of sequence. (From de Leval MR, 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).

Figure 2. VLAD for series of 393 operations performed by single surgeon expressed as expected minus observed deaths according to operation type, age, left ventricular function, urgency of operation, and previous cardiac surgery.⁸ Prediction intervals represent distribution of outcomes at end of series if all observed variation is due to chance.

Figure 3. VLAD for group of 6 surgeons. Expected mortality was calculated according to Parsonnet model. Notice that after run of deaths (bold line), 1 surgeon stopped operating after discovery of a cortical visual handicap.

3 would surely, if seen in real time, have prompted a question. What statistical test might have been used is a different issue; not all instances of deteriorating performance need statistical proof or merit statistical analysis. When an institution suspects deteriorating performance, it needs to be demonstrated explicitly for fair and open analysis and discussion.

Our second example deals with results of a single surgeon during 1990 to 1994, well before the issues surrounding performance monitoring heated up (Figure 4). The surgeon was believed by colleagues to be underperforming, but believed himself to have acceptable results given the case mix undertaken, and indeed he was not a man to shy away from risk. He defended himself vigorously against what he regarded as an attempt at unfair dismissal and took his employing hospital to court. His colleagues had the sorry

task of explaining Figure 4 to lawyers. Earlier explicit display of performance in an intuitively obvious form might have saved some lives, as well as the surgeon's reputation.

We cannot assume that a human being will recognize the first decline in performance, stop, and seek help. On the contrary, a surgeon who cannot live with risk, who cannot get back in the saddle after being thrown, is probably not in the right job. It is a fact of life, albeit unpalatable, that doctors need external agencies to regulate their performance.

VLAD charts are gaining in popularity because they are intuitive and can reveal trends meriting inspection and discussion. We have given some examples where display of the results is itself enough to lead to an explanation and a solution. We are left, however, with the problem of how to decide what is acceptable, what is questionably acceptable, and how to determine at what point results have become unacceptably bad. The debate about whether we can or should construct control limits on the charts is an important

one,^{3,4,6} but it is also important to remember that displaying sequential risk-adjusted data does not create a problem, although it may reveal one. We have always had to decide how to declare what is “significantly” worse than an acceptable standard. It is inherent in the problem that an alert (to use de Leval’s term¹) must signal before the conventional level of scientific proof required to test a scientific hypothesis; if we allow events to run their course until a conventional level of significance is reached (such as $P = .05$), many lives will have been lost.

Whenever results of treatment are collected, it would be sensible to ask, “For what purpose?” Current pressure in the United Kingdom to collect and make available mortality data is for the early recognition of episodes of underperformance, whether from the surgeon’s skill, the institution’s systems, or any other cause. The SCTS has decided that it will present non-risk-adjusted mortality data with 99.99% confidence intervals. This means that only once in 10,000 instances would a run of deaths occurring by chance (that is to say, common-cause variation⁴) be unfairly attributed to a surgeon experiencing a run of “bad luck.” Of course, the corollary is that true problems may not be detected in time to avert disaster—and yet one might have thought that was the original purpose of monitoring. In defense of this decision, the SCTS expresses the view that those on the inside will already have discovered that performance is slipping, which is a way of saying that the open publication of results is window dressing for public consumption, and meanwhile, surgeons will look after their own affairs. The SCTS is of the view that real problems will be picked up long before their proposed test becomes positive, so the public will never see a surgeon’s results cross the $P = .0001$ threshold. Let us hope so for the patients’ sake (for that represents a lot of deaths), the surgeon’s sake (for that will be a mighty fall), and for the reputation of the surgical profession, which in the United Kingdom has suffered enough suspicion and criticism in recent years. “Trust me—I’m a doctor” is no longer sufficient reassurance.

For those who remain uncomfortable without P values, formal statistical inference derived from sequential mortality data requires great care.⁶ Certainly VLAD charts should not be used as a method for formal hypothesis testing. Whatever methods are used for such a purpose, they must take into account repeated testing and the fact that successive hypothesis tests are carried out on data sets that overlap. Case-mix correction considerably complicates hypothesis testing. For example, it may be mathematically convenient to test the hypothesis that there has been a uniform inflation of all risks by the same factor—the assumption in the control limit methods described by Rogers

Figure 4. Sequential risk-adjusted graph (VLAD method) of single surgeon and results from 1990 through 1994. Year after year, this surgeon had risk-adjusted mortality figures worse than those expected from Parsonnet model.

and colleagues.⁴ Whether this reflects the realities of what happens if a surgeon’s performance declines is dubious, because one might expect a disproportionately high increase in mortality for technically challenging rather than routine procedures. Further, we have noted that surgeons with overall excellent results have had runs of as many as 50 to 100 operations in which their results fell below those expected according to a risk model, even for the now too-forgiving Parsonnet model.

We reiterate all the caveats of Rogers and colleagues,⁴ in particular that VLAD was intended as a means of data display and with appropriate control limits might serve as a statistical “ready reckoner,” but it should not be used as a hypothesis testing method.

References

1. de Leval MR, 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;107:914-23.
2. 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.
3. Grunkemeier GL, Wu YX, Furnary AP. Cumulative sum techniques for assessing surgical results. *Ann Thorac Surg.* 2003;76:663-7.
4. Rogers CA, Reeves BC, Caputo M, Ganesh JS, Bonser RS, Angelini GD. Control chart methods for monitoring cardiac surgical performance and their interpretation. *J Thorac Cardiovasc Surg.* 2004;128:811-9.
5. Keogh B, Kinsman R. National adult cardiac surgical database report 2000-2001. London: Society of Cardiothoracic Surgeon of Great Britain and Ireland; 2002.
6. Sherlaw-Johnson C, Lovegrove J, Treasure T, Gallivan S. Likely variations in perioperative mortality associated with cardiac surgery: when does high mortality reflect bad practice? *Heart.* 2000;84:79-82.
7. Parsonnet V, Dean D, Bernstein AD. A method of uniform stratification of risk for evaluating the results of surgery in acquired adult heart disease. *Circulation.* 1989;79(6 Pt 2):13-12.
8. Lovegrove J, Sherlaw-Johnson C, Valencia O, et al. Monitoring the performance of cardiac surgeons. *J Opl Res Soc.* 1999;50:684-9.