Monitoring surgical performance

Eugene H. Blackstone, MD

See related articles on pages
 811, 820, 823, and 907.

If surgical performance—often measured by postoperative outcome of initial hospital stay—is monitored at all, the most common means is by risk-adjusted annual or semiannual audit. Observed occurrence of outcome measures (eg, in-hospital death and complications) as a proportion of cases performed is compared with expected performance using, for example, the Society of Thoracic Surgeons' regression equations¹ or EuroSCORE,² which account for many aspects of case mix. Sometimes observed (O) and expected (E) proportions are subtracted, sometimes divided (O/E ratio)³; sometimes confidence limits of these comparisons are provided, and occasionally *P* values are given.

Is this periodic, widespread, but rather coarse monitoring of surgical performance sufficient?

CUSUMs—What Are They and Why Should We Care?

During the past year, several manuscripts have been submitted to this *Journal* and to *The Annals of Thoracic Surgery* that argue for monitoring on an individual case-after-case fashion (sequential monitoring) using the ever-expanding suite of statistical quality control techniques.⁴ The most commonly used belong to the family of cumulative sum (CUSUM) charts.⁵ These charts purport to provide early identification of deviation from a performance standard. Their forte is identifying subtle, slow, but sustained degradation in a system thought to be in control.⁴ Because early, reliable warning should be good for both patients and surgeons, the editors decided that a “Statistics for the Rest of Us” tutorial on CUSUMs would be valuable for readers.

No sooner had this decision been made than (1) Gary Grunkemeier's tutorial on CUSUMs appeared in *The Annals*,⁶ and (2) we received a tutorial by Rogers and colleagues⁷ from Bristol. We determined that the latter would become the centerpiece of an educational package on monitoring surgical performance, along with invited commentaries from Tom Treasure and his group from Guy's Hospital and the Clinical Operational Research Unit, Department of Mathematics, University College London, and David Spiegelhalter from the Medical Research Council Biostatistics Unit, Cambridge.

We trust that this material, together with Grunkemeier's *Annals* presentation,⁶ will provide you, the reader, with a comprehensive idea of the “state of the art” in surgical performance monitoring. We have purposely retained controversy, even potentially inflammatory statements, because the field of quality monitoring in medicine, and even in industrial settings, is still evolving.

Some Things to Look for

As you read the tutorial and accompanying commentaries, look for a few facts, ideas, differences of opinion, and areas of incomplete knowledge.

From the Section of Clinical Research, Department of Thoracic and Cardiovascular Surgery and Department of Biostatistics and Epidemiology, The Cleveland Clinic Foundation, Cleveland, Ohio.

Received for publication Feb 24, 2004; accepted for publication March 5, 2004.

Address for reprints: Eugene H. Blackstone, MD, Department of Thoracic and Cardiovascular Surgery, The Cleveland Clinic Foundation, 9500 Euclid Avenue, Desk F24, Cleveland, OH 44195 (E-mail: blackse@ccf.org).

J Thorac Cardiovasc Surg 2004;128:807-10
 0022-5223/\$30.00

Copyright © 2004 by The American Association for Thoracic Surgery

doi:10.1016/j.jtcvs.2004.03.022

History

Although Rogers and colleagues⁷ claim that Williams and colleagues⁸ first proposed using CUSUMs in a medical context, this is untrue. CUSUM techniques have been used quite effectively in medicine for at least 35 years, and control charts have been used for at least 50 years.⁹ Initially they were used mostly to monitor quality of clinical chemistry laboratory measurements,^{10,11} but in 1977 *The New England Journal of Medicine* published Herbert Wohl's article, "The CUSUM Plot: Its Utility in the Analysis of Clinical Data."¹² Wohl illustrated use of CUSUM charts for detecting subtle body temperature changes in patients being treated for sepsis. I mention this article not just because of the prestigious journal in which it appeared, but to contrast *continuous* monitoring with that of *single outcomes* of discrete patients. Because temperature measurements can be recorded continuously—just as can thickness of a rolled sheet of steel—sustained temporal trends can be detected quickly. But as Rogers and colleagues⁷ point out, CUSUM techniques may require years of patients in low-volume settings to detect performance problems measured as binary outcomes.

Although de Leval and colleagues¹³ are credited with introducing non-risk-adjusted CUSUM charts to cardiac surgeons, as pointed out by Treasure and colleagues in their commentary, often missed are two other ideas introduced in their report. First, they dealt with the problem, dismissed by Rogers and colleagues,⁷ that traditional CUSUM techniques have no memory loss. If continuous monitoring of a program is being suggested, how long is it necessary to remember and equally weight results of the past? de Leval and colleagues suggested using an exponential memory loss, called "exponentially weighted moving average" (EWMA) charts by Spiegelhalter in his commentary. Second, they introduced a form of risk adjustment that may still be valuable. They simply superimposed on their CUSUM results of *observed* outcome an *expected* CUSUM outcome calculated from an external¹⁴ risk-adjustment equation (their Figure 4).¹³

Performance Improvement

Both Rogers and colleagues⁷ and the commentators use two phrases that may be unfamiliar to readers: *common-cause* variation and *special-cause* variation. Common-cause variation is the natural fluctuation of performance measures that results from multiple factors underlying any complex process, such as health care, that is considered to be in control. James Reason,¹⁵ in discussing human error, and W. Edwards Demming,¹⁶ in discussing industrial processes, emphasize that nearly all improvement in results or product come from reducing common-cause variation (Reason calls it the "blunt end"). Special-cause variation is fluctuation in results that are attributed to those aspects of the process over which there is presumed to be some extrinsic influence, such

as that of the surgeon. Reason argues that improvement in this source of variation at the "sharp end" of patient care delivery is most effective in a non-culpable atmosphere, because things are rarely as simple as a single individual to blame.¹⁷ (An alternative is to institute mechanisms to insulate the process from blunt-end systems.)

Performance Measures

Not discussed in depth by Rogers and colleagues⁷ nor by the commentators are appropriate measures of performance. It is possible that several hospital outcomes should be simultaneously monitored, and this is what Spiegelhalter has termed "multiplicity." Silber and colleagues^{18,19} have emphasized the difficulties of selecting outcome measures that reflect controllable variation and are not confounded by patient factors. The fact that risk-adjustment methods are advocated by all the discussants indicates that the outcomes selected for monitoring are thought to be strongly confounded by patient and disease characteristics. Unfortunately, risk adjustment tends to be particularly incomplete when there are rare or multiple measured, unmeasured, or unevaluated risk factors present,²⁰ so the search for adequate unconfounded quality measures should go on.

Response Speed

We have already alluded to the difference in response speed to underlying trends when a continuous variable is monitored, such as temperature (the kind of things often measured in industrial quality control) as opposed to one value from an entire operative result. The discussants have focused on boundary crossing methods to detect these trends. Yet CUSUM charts, in contrast to a number of other kinds of chart, are considered most valuable for detecting a change in slope.^{4,21}

Some have hyped CUSUMs as instantaneous warning systems for undesired outcomes. Untrue. I agree with Lim²² that we need more sensitive and responsive warning systems, but if mortality is the performance measure, do not expect it to provide instantaneous warning.

Comparison of Surgical Programs

The Society of Thoracic Surgeons, its European counterparts, and governmental agencies compare not only individual surgeon performance but institutional performance as well. Most models for monitoring have been constructed without taking into account institutions and surgeons.¹⁻³ Subsequently, models are applied on an institution or surgeon basis. Tekkis and colleagues,²³ in the setting of gastroesophageal cancer surgery, use hierarchical (mixed) modeling that permits direct assessment of institutional performance while simultaneously modeling underlying risk factors. This approach is explained in an accessible tutorial by Christiansen and Morris,²⁴ as cited by Spiegelhalter. It is an attempt to model simultaneously both special-cause and

common-cause variation. Such an approach, even if performed only periodically, has considerable merit.

Simplicity and Intuitiveness

Consistent with other recent developments in statistical quality control, Shewhart's original sketch of a quality control chart at Bell Telephone Laboratories on May 16, 1924, was simple and intuitive.²⁵ Figure 1 represents the kind of simplicity originally envisioned for control charts. An in-control process marches horizontally down the centerline, staying out of areas of alarm. The underlying mathematics of Rogers and colleagues' hypothesis testing approach (verified by the mathematics in their Appendix)⁷ (1) do not require that in-control processes march down either a centerline or a fixed slope corresponding to the in-control observations and (2) do require an in-control process to march toward an acceptance boundary. Neither behavior is simple or intuitive. Similarly, I find Grunkemeier's bullet-shaped prediction limit approach (which some might mistake for control limits) equally nonintuitive, because the limits present moving targets dependent on number rather than standards of performance.⁶ The most intuitive chart to my eye is the observed minus expected chart for which Rogers and colleagues⁷ do not display boundary lines.

Multiple Testing

All discussants argue (from either assumptions or particular schools of statistical thought) that continual testing is in some sense subject to the multiple comparison problem, and one's interpretation must be altered by how often the data are evaluated. Some statisticians vigorously defend this view; others vigorously hold that the multiple comparison problem is not applicable to the quality control setting. They point to the continuous monitoring of thickness of sheet Mylar, for example, as it is produced by DuPont, as being equivalent to an infinite number of checks for which adjustment for multiple comparisons would result quickly in wide, unusable limits. Instead, natural variation in thickness (common-cause variation) is well quantified or specified, and deviations from specified tolerances signal an alarm. Slow trends away from specified thickness are hard to detect unless the fluctuations are integrated across time, hence the need for the continuous CUSUM in such a setting to detect change in slope. There are no considerations of sample size, multiple comparisons, or hypothesis testing! Indeed, in the de Leval article with Spiegelhalter,¹³ the authors argue that "the CUSUM procedure... avoids the well-known problem associated with repeated significance testing," based on the work of several statisticians.^{8,26,27}

I am not sure what to believe, frankly, nor do I think this issue will be soon resolved. However, Storey's work at Stanford University on false discovery rates²⁸ and Aylin and colleagues' work²⁹ seem to be promising and fresh approaches to this problem, as noted by Spiegelhalter.

Figure 1. Format for simple, intuitive control chart, such as originally envisioned by Shewhart.²⁵ Along the vertical axis is quantity (outcome) being monitored and on the horizontal axis either time or sequence number of each operation. Middle horizontal line is positioned at a value representing an "in-control" system. Presumably, random (common-cause) variation stays within alert lines. When graphed results stray outside these lines, particularly above the lines if increasing value on the vertical axis corresponds to poorer performance, investigation of cause for this behavior is initiated. Alternatively, for a CUSUM approach that integrates fluctuations, an upward-trending slope may earlier signal a system that appears to be headed toward being out of control.

What's Next?

CUSUMs are not the end of the road for techniques that may be useful for surgical performance monitoring. My digital signal processing background conjures up visions of applying sophisticated pattern recognition techniques, such as wavelet kernels, to identify underlying trends and transients. Might optimal statistical outlier identification methods be an alternative approach? Algorithmic technologies may yield yet other methods.³⁰

The Shocker

Most readers of the tutorial and commentaries will be surgeons or physicians involved in health care delivery. Particularly in a litigious society, health care workers want to be given the benefit of the doubt. On the other hand, when the tables are turned and you become the patient, would you not find it shocking that the Society of Cardiothoracic Surgeons of Great Britain and Ireland interprets "benefit of the doubt" to mean 9999:1 odds of adverse outcomes being attributable to chance alone before calling those results into question? Protecting our own reputations versus protecting our patients' lives involves a delicate and sensitive balance between being an alarmist and being insensitive.²² Tipping the balance decidedly in favor of our own interests versus those of our patients can only toss fuel onto a fire that is burning up the public's confidence in the medical profession. On the other hand, monitoring programs that fail to recognize that systems, not individuals at the sharp end of

the process, should be the prime targets for quality improvements will continue to concentrate on sniffing out “bad eggs.” They address the proverbial speck in the eye rather than first removing the plank in the eye of the blunt end of medical care.

References

- Shroyer AL, Plomondon ME, Grover FL, Edwards FH. The 1996 coronary artery bypass risk model: the Society of Thoracic Surgeons Adult Cardiac National Database. *Ann Thorac Surg.* 1999;67:1205-8.
- Nashef SA, Roques F, Michel P, Gauducheau E, Lemeshow S, Salamon R. European system for cardiac operative risk evaluation (EuroSCORE). *Eur J Cardiothorac Surg.* 1999;16:9-13.
- Hannan EL, Kumar D, Racz M, Siu AL, Chassin MR. New York State's Cardiac Surgery Reporting System: four years later. *Ann Thorac Surg.* 1994;58:1852-7.
- Montgomery DC. Introduction to statistical quality control. 4th ed. Hoboken (NJ): John Wiley & Sons; 2000.
- Page E. Continuous inspection schemes. *Biometrika.* 1954;41:100-14.
- Grunkemeier GL, Wu YX, Furnary AP. Cumulative sum techniques for assessing surgical results. *Ann Thorac Surg.* 2003;76:663-7.
- 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.
- Williams SM, Parry BR, Schlup MM. Quality control: an application of the cusum. *BMJ.* 1992;304:1359-61.
- Levey S, Jennings ER. The use of control charts in the clinical chemical laboratory. *Am J Clin Pathol.* 1950;20:1059-66.
- Riddick JH, Giddings NW. Computerized preparation of average CUSUM charts for clinical chemistry. *Clin Biochem.* 1971;4:156-61.
- Westgard JO, Groth T, Aronsson T, de Verdier CH. Combined Shewhart-cusum control chart for improved quality control in clinical chemistry. *Clin Chem.* 1977;23:1881-7.
- Wohl H. The cusum plot: its utility in the analysis of clinical data. *N Engl J Med.* 1977;296:1044-5.
- 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.
- Kirklin JW, Blackstone EH, Tchervenkov CI, Castaneda AR. Clinical outcomes after the arterial switch operation for transposition. Patient, support, procedural, and institutional risk factors. Congenital Heart Surgeons Society. *Circulation.* 1992;86:1501-15.
- Reason J. Human error. Cambridge, UK: Cambridge University Press; 1999.
- Deming WE. Out of the crisis. Cambridge (MA): MIT Press; 1986.
- Wigglesworth EC. A teaching model of injury causation and a guide for selecting countermeasures. *Occup Psychol.* 1972;46:69-78.
- Silber JH, Rosenbaum PR, Schwartz JS, Ross RN, Williams SV. Evaluation of the complication rate as a measure of quality of care in coronary artery bypass graft surgery. *JAMA.* 1995;274:317-23.
- Silber JH, Williams SV, Krakauer H, Schwartz JS. Hospital and patient characteristics associated with death after surgery. A study of adverse occurrence and failure to rescue. *Med Care.* 1992;30:615-29.
- Sergeant P, Blackstone E, Meyns B. Can the outcome of coronary bypass grafting be predicted reliably? *Eur J Cardiothorac Surg.* 1997;11:2-9.
- Koning AJ. CUSUM charts for preliminary analysis of individual observations. *J Qual Technol.* 2000;32:122-32.
- Lim TO. Statistical process control tools for monitoring clinical performance. *Int J Qual Health Care.* 2003;15:3-4.
- Tekkis PP, McCulloch P, Steger AC, Benjamin IS, Poloniecki JD. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *BMJ.* 2003;326:786-8.
- Christiansen CL, Morris CN. Improving the statistical approach to health care provider profiling. *Ann Intern Med.* 1997;127:764-8.
- Shewhart WA. Economic control of quality of manufactured product. Princeton (NJ): Van Nostrand Reinhold; 1931.
- McPherson K. Statistics: the problem of examining accumulating data more than once. *N Engl J Med.* 1974;290:501-2.
- Kenett R, Pollak M. On sequential detection of a shift in the probability of a rare event. *J Am Stat Assoc.* 1983;78:389-95.
- Storey JD. A direct approach to false discovery rates. *J R Statist Soc B.* 2002;64:479-98.
- Aylin P, Best N, Bottle A, Marshall C. Following Shipman: a pilot system for monitoring mortality rates in primary care. *Lancet.* 2003;362:485-91.
- Breiman L. Statistical modeling: the two cultures. *Stat Sci.* 2001;16:199-231.

The Journal of Thoracic and Cardiovascular Surgery Conflict of Interest Policy

To assure fairness to authors submitting work for consideration in *The Journal of Thoracic and Cardiovascular Surgery*, a mechanism exists for managing conflicts of interest. The editor and each of the section editors complete a “Conflict of Interest” form that identifies any and all relationships with commercial and other academic entities. When the editor has a potential conflict because of a relationship with another entity or author, the editor appoints an alternate editor from among the section editors or editorial board members who assumes the entire responsibility for final decisions on the manuscript in question. The editor does not read the reviews that are submitted nor engage in discussing the manuscript prior to the final decision. When the conflict of interest involves a section editor, a “guest section editor” is appointed who fills the role normally played by the conflicted section editor. All members of the editorial board and reviewers are asked to indicate any conflict of interest when they agree to review a manuscript.