Monitoring clinical performance: A commentary

D. J. Spiegelhalter, PhD

See related article on pages 807, 811, 823, and 907.

I am glad to have the opportunity to comment on the article by Rogers and colleagues,¹ which contains an excellent discussion of both the potential benefits and hazards of using formal statistical monitoring procedures for clinical data. I particularly welcome the manner in which the authors warn about the possible misinterpretations of the graphs and their recommendation to view multiple complementary charts.

It would be easy to scoff at the growing enthusiasm and spirit of discovery concerning such monitoring procedures, given that Shewhart control charts were being recommended in an industrial context more than 70 years ago,² and sequential probability ratio tests (SPRT) were developed independently by Wald³ in the United States and Barnard⁴ in the United Kingdom in the early 1940s. But the recent industrial cumulative sum (CUSUM) literature, for example Hawkins and Olwell,⁵ shows that many research issues of crucial importance appear to be outstanding, for example, risk adjustment, allowance for overdispersion, multiple systems, and so on. Increased demands for quality assurance in complex medical systems are fueling strong interest in good methodology,⁶ and previously exotic phrases such as “risk-adjusted CUSUMs” are becoming standard language. This is a particularly exciting period in collaborative research.

My comments concern seven main aspects that need to be considered when designing any monitoring system and focus on what I believe to be currently important issues. The term *system* is used to imply a process that involves more than, say, a single surgeon analyzing results of a single procedure; rather, it involves some monitoring body given responsibility for quality assurance of a number of different outcomes.

1. *Unit being monitored.* Traditionally, discussion of statistical process control methods has focused on a single “unit” being monitored, whether an industrial process or a surgeon. But an organization charged with quality assurance may be monitoring many such units, and the properties of any system need to be evaluated accordingly. This multiplicity issue may lead one to consider a composite model of all the units, for example, a hierarchical model leading to “shrunk estimates,”⁷ or an alternative to standard type I error rates, such as estimation of false discovery rates,⁸ essentially the predictive value of a positive signal.
2. *Event being monitored.* Again, the traditional focus has been on a single series, whereas in practice a health care system is likely to be monitoring many, perhaps hundreds, of indicators. Some selection of indicators, or amalgamation into composites, may be possible, but this additional source of multiplicity also needs to be investigated when characterizing the properties of any system.
3. *Risk adjustment of outcomes.* Rogers and colleagues¹ state the importance of risk adjustment, but also recognize that it is always inadequate. One response is to be more flexible in the “target” performance by allowing a certain degree of variability due to the effects of many small, unobserved risk factors. This leads naturally to allowance for overdispersion.⁹
4. *Choice of summary statistic for monitoring.* As Rogers and colleagues¹ emphasize, it is inappropriate to try to select a single summary statistic to monitor. Different choices have different strengths in terms of interpretability and statistical properties. For example, Figure 1 shows the experience of serial murderer Dr Harold Shipman (data provided by Baker¹⁰ and previously

From the Medical Research Council Biostatistics Unit, Institute of Public Health, Cambridge, United Kingdom.

Received for publication Feb 20, 2004; accepted for publication March 4, 2004.

Address for reprints David J. Spiegelhalter, PhD, Medical Research Council, Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CR2 2SR, United Kingdom (E-mail: david.spiegelhalter@mrc-bsu.cam.ac.uk).

J Thorac Cardiovasc Surg 2004;128:820-2
0022-5223/\$30.00

Copyright © 2004 by The American Association for Thoracic Surgery

doi:10.1016/j.jtcvs.2004.03.024

Figure 1. Sequential monitoring of death certificates signed by Dr Harold Shipman. Top left: Observed and expected number of deaths, with expected number based on experience of local colleagues. **Bottom left:** Cumulative observed minus expected deaths (cumulative excess mortality). **Top right:** Shewhart charts of standardized excess deaths. Shewhart chart limits (dot-dash-dot lines) are based on 3 SE difference in standardized excess mortality. (Note: Control limits have been adjusted for overdispersion multiplier of 3.42 to allow for inadequate risk adjustment, per Aylin and colleagues.⁹) **Bottom right:** Log-likelihood ratio risk-adjusted CUSUM (RA-CUSUM). RA-CUSUM is based on log-likelihood ratio's ability to detect doubling of risk (dot-dash-dot line). In theory, this procedure might have detected excess mortality in Shipman's experience in 1984, after which he went on to kill at least 150 more people. In practice, however, a facility for monitoring mortality of a general practitioner would have been impossible, because no linkage between records was made.

analyzed by Spiegelhalter and colleagues.¹¹) Control limits have been adjusted for the overdispersion factor found by Aylin and colleagues.⁹ By plotting both the observed and expected deaths (*top left*), we can see that some of the excess in later years resulted from a decline in expected deaths (in part because Dr Shipman had already murdered a substantial proportion of his older patients). The Shewhart chart (*top right*) shows when individual years were out of control, the cumulative excess mortality chart (*bottom left*, also known as VLAD or CRAM), is easy to interpret, although the risk-adjusted CUSUM (RA-CUSUM; *bottom right*) formally provides a most powerful test. There seems no reason to select a single chart to display; the RA-CUSUM may be used as a formal trigger of an alarm, but the other plots aid in interpretation. My personal preference is now for the RA-CUSUM¹² rather than the SPRT. As Rogers and colleagues¹ point out, the RA-CUSUM has the same steps as the SPRT, but is constrained to lie above 0. This means that it cannot build up credit and so retains

sensitivity, but formally has type I error of 1 and type II error of 0, because it is guaranteed to eventually reject the null.¹³

5. *Role of discounting past experience.* It seems reasonable to discount past experience to some extent, because this leads to an attractive emphasis on *current* performance and so avoids having to specify a fixed start time for the monitoring process. Discounted methods lead naturally to use of exponentially weighted moving average (EWMA) estimates, as used by de Leval and associates,¹⁴ which are fine as sequential estimates. It would seem natural to investigate developments such as risk-adjusted EWMA and exponentially weighted RA-CUSUMs.
6. *Selection of thresholds for action.* Rogers and colleagues¹ set their thresholds on the attractive and familiar basis of type I and type II error rates. However, this is not appropriate for risk-adjusted RA-CUSUMs, and the numerous allowances for multiplicity described previously also play havoc with any attempt to specify error rates. Perhaps a more prag-

matic approach would be to investigate a range of potential thresholds in terms of their performance both on simulated data and also on real past data. Aylin and colleagues⁹ have adopted this approach for RA-CUSUMs, reporting both sensitivity and false discovery rates across a limited period.

7. *Actions to be taken.* Strictly speaking, it is impossible to set thresholds unless one knows what actions are to be taken in response to their being crossed, and hence the design of a monitoring system must be intimately driven by its context and intended use. Any response should be in a prespecified and staged manner, with the first step being to check the data carefully! It is vital that these actions be considered at the design stage to avoid a well-designed statistical monitoring system becoming discredited through inappropriate use.

I feel fortunate to be working in an area in which good methodology, founded on strong traditional principles but updated in response to modern demands, can be applied to important problems. The article by Rogers and colleagues¹ is a valuable contribution to this movement.

References

1. 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.
2. Shewhart WA. Economic control of quality of manufactured product. Princeton (NJ): Van Nostrand Reinhold, 1931.
3. Wald A. Sequential tests of hypotheses. *Ann Math Statist.* 1945;16:117-86.
4. Barnard GA. Sequential tests of statistical hypotheses. *J R Statist Soc.* 1946;8(Suppl):1-26.
5. Hawkins DM, Olwell DH. Cumulative sum charts and charting for quality improvement. New York: Springer; 1998.
6. Sonesson C, Bock D. A review and discussion of statistical issues in public health monitoring. *J R Statist Soc Ser C.* 2003;166:5-21.
7. Christiansen C, Morris C. Improving the statistical approach to health care provider profiling. *Ann Intern Med.* 1997;127:764-8.
8. Storey JD. A direct approach to false discovery rates. *J R Statist Soc Ser B.* 2002;64:479-98.
9. Aylin P, Best NG, Bottle A, Marshall C. Following Shipman: a pilot system for monitoring mortality rates in primary care. *Lancet.* 2003;362:485-91.
10. Baker R. Harold Shipman's clinical practice 1974-1998: a review commissioned by the chief medical officer. London: The Stationery Office; 2001.
11. Spiegelhalter DJ, 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.
12. Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics.* 2000;1:441-52.
13. Grigg O, Farewell VT, Spiegelhalter DJ. A comparison of approaches to sequential monitoring of risk-adjusted health outcome. *Stat Methods Med Res.* 2003;12:147-70.
14. de Leval MR, Francois K, Bull K, Brawn W, Spiegelhalter DJ. Analysis of a cluster of surgical failures: application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg.* 1994;107:914-24.

JTCVS On-Line Manuscript Submission and Review

Please visit <http://www.editorialmanager.com/jtcvs/>

Effective September 15, 2001, authors and reviewers may submit manuscripts and reviews electronically via Editorial Manager, our new Web-based system with full electronic submission, review, and status update capabilities.

As we move from paper to electronic submissions, the Editorial Office will make proxy submissions of all manuscripts accompanied by a diskette containing the electronic files of the text, tables, and figures. Editors, authors, and reviewers will receive automatic e-mails when significant events occur.

We strongly encourage all authors and reviewers to use Editorial Manager. Although we will continue to accommodate the submission of paper manuscripts for some months, our goal is to be completely electronic within 9 to 12 months.

All individuals currently in our database for whom we have e-mail addresses will receive via e-mail a system-assigned username and password that can be used to log in to the system without prior registration. All those not receiving the e-mail must register the first time they use the system.

As with any broad systemic change, the conversion to the new system will take some time to complete. We ask your patience as we replace our in-office database with the new system. We also encourage you to take advantage of the speed and efficiency that the new system will provide for us all: editor, author, reviewer, and publisher.