International Journal for Quality in Health Care 2003; **Volume 15, Number 1**: pp. 7–13

Risk-adjusted sequential probability ratio tests: applications to Bristol, Shipman and adult cardiac surgery

DAVID SPIEGELHALTER[1], OLIVIA GRIGG[1], ROBIN KINSMAN[2] AND TOM TREASURE[3]

[1]MRC Biostatistics Unit, Institute of Public Health, Cambridge, [2]Dendrite Clinical Systems, Henley-on-Thames, Oxfordshire, [3]Thoracic Surgery, Guy's and St Thomas Hospitals, London, UK

Abstract

Objective. To investigate the use of the risk-adjusted sequential probability ratio test in monitoring the cumulative occurrence of adverse clinical outcomes.

Design. Retrospective analysis of three longitudinal datasets.

Subjects. Patients aged 65 years and over under the care of Harold Shipman between 1979 and 1997, patients under 1 year of age undergoing paediatric heart surgery in Bristol Royal Infirmary between 1984 and 1995, adult patients receiving cardiac surgery from a team of cardiac surgeons in London, UK.

Main outcome measure. Annual and 30-day mortality rates.

Results. Using reasonable boundaries, the procedure could have indicated an 'alarm' in Bristol after publication of the 1991 Cardiac Surgical Register, and in 1985 or 1997 for Harold Shipman depending on the data source and the comparator. The cardiac surgeons showed no significant deviation from expected performance.

Conclusions. The risk-adjusted sequential probability test is simple to implement, can be applied in a variety of contexts, and might have been useful to detect specific instances of past divergent performance. The use of this and related techniques deserves further attention in the context of prospectively monitoring adverse clinical outcomes.

Keywords: adverse clinical outcomes, general practitioners, monitoring, mortality, paediatric and adult cardiac surgery, risk-adjustment, sequential probability ratio tests

The need for systems to monitor clinical performance has been brought into particular focus by the report of the Bristol Royal Infirmary Inquiry [1] and the finding that general practitioner Harold Shipman murdered over 200 of his patients [2]. Rarer adverse events may require cumulative monitoring over time rather than, for example, examination of annual data. Historical industrial quality control procedures have been recommended, such as Shewhart control charts and cumulative sum (CUSUM) techniques [3,4]. The medical context does, however, add an additional complexity in the need to adjust for case-mix in an attempt to avoid clinicians or trusts being unfairly penalized for treating higher-risk patients. A common suggestion is to plot the accumulating difference between the observed number of adverse events, and the number expected according to an established risk-adjustment procedure [5,6]. If concerned with mortality, for example, this means the cumulative 'excess deaths' (or conversely 'lives saved') can be monitored [7].

The problem then arises of setting appropriate 'thresholds' on such plots to indicate the need for further scrutiny. These can be set for a single time point using standard methods for confidence intervals [8]. However, this does not allow for the well-known problem of repeated testing in which a true null hypothesis is certain to be eventually rejected [9], which in this context could correspond to an unreasonable number of false accusations of poor performance. The 'risk-adjusted CUSUM' has been suggested as a technique for dealing with both risk-adjustment and sequential testing, but setting appropriate thresholds is not straightforward [10].

In this paper we investigate a risk-adjusted version of the classic sequential probability ratio test (SPRT) that was developed for quality control of military supplies in the

Address reprint requests to Olivia Grigg, MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, UK. E-mail: olivia.grigg@mrc-bsu.cam.ac.uk

Second World War: this is shown to take the form of a simple adaptation of a cumulative 'observed—expected' plot with horizontal thresholds. Examples are provided in three monitoring contexts: annual surgical mortality (paediatric cardiac surgery in Bristol), adverse events in a population (Harold Shipman's practice) and individual operations (cardiac surgery by a group of surgeons). We conclude that the risk-adjusted SPRT is a simple technique that could be useful within a clinical monitoring system.

Materials and methods

Formal statistical methods for sequential analysis were developed in 1943 independently by Barnard in the UK and Wald in the US [11,12]. The SPRT is the most powerful method for discriminating between two hypotheses [12], and was recommended well over 40 years ago in a medical context for clinical trials and clinical experiments [13,14].

The formal procedure has three components: a running test statistic, thresholds for the statistic that determine statistical significance, and actions to be taken on crossing a threshold.

The test statistic

Suppose we have two hypotheses: a null hypothesis H_0 corresponding to performance as expected, and H_1 to a level of performance deemed importantly divergent. Wald showed that the most powerful sequential comparison based on an accumulating set of data takes the form of a running 'log-likelihood ratio' (LLR) [12], which is increased or decreased after each observation by a quantity depending on the event observed and, in its risk-adjusted form, the expected outcome if that event were performance 'as normal', i.e. under H_0. This procedure is described in the Appendix. Here, we have set our example charts to detect a doubling in the relevant process parameter, purely for illustration. In practice, one should set a chart to detect the lowest of the range of values deemed unacceptable.

Thresholds

Wald showed that sampling should continue if the LLR lies between two thresholds denoted a and b: when LLR exceeds b, stop and reject H_0 in favour of H_1, and vice versa when LLR is less than a [12]. Thus, the boundaries take the form of horizontal lines. Using the traditional language of statistical hypothesis testing, let

α = probability of eventually rejecting H_0 when it is true (Type I error)

β = probability of eventually rejecting H_1 when it is true (Type II error)

If we are willing to choose values for α and β, we can use the following equations, developed by Wald [12], to calculate (approximately) appropriate thresholds a and b

$$a = \log[\beta/(1-\alpha)]$$

$$b = \log[(1-\beta)/\alpha]$$

The sizes of α and β should reflect the relative 'costs' of making the two types of error. For example, if we wish to avoid falsely identifying an adequate surgeon as 'higher-risk', then α should be made very small, whereas if we consider it a very serious mistake to miss a poorly performing surgeon, then β should be made very small. We have adopted a convention of equal α and β, with illustrative examples given in Table 1 of the Appendix.

Instead of choosing a single value for α and β, a set of horizontal lines can be drawn on the chart to indicate different degrees of urgency: for example, a single centre might use $\alpha = \beta = 0.1$ as an 'alert' threshold and $\alpha = \beta = 0.01$ for 'alarm'. When monitoring many individuals or centres more stringent boundaries may be appropriate because of the many comparisons being made: of 10 centres performing normally, we would expect one to cross the 'alert' boundary by chance alone. A Bonferroni-like adjustment might suggest, when monitoring n individuals or institutions, using $0.10/n$ and $0.01/n$ for 'alert' and 'alarm' respectively.

Action

Strictly speaking, we should pre-specify values for α and β and stop monitoring once either threshold has been crossed. An alternative is to restart the procedure when, say, we cross the lower boundary and so are confident there is no increase in mortality. This has the advantage that it is not possible to build up excessive 'credit' and so gains sensitivity to changes in performance, but also has the disadvantage that the strict interpretation of α and β is lost. However, as we take the view that any such procedure is only an aid to clinical monitoring, we shall allow for restarts when crossing the 'reject H_1' boundary in our examples. However, when divergent performance has been detected ('reject H_0') we shall not assume a restart takes place.

Three datasets are used for illustration.

Bristol

As part of the Bristol Royal Infirmary Inquiry [1], annual mortality rates for open-heart surgery on children under 1 year of age were made available from the Cardiac Surgical Register between 1985 and 1995, and from Hospital Episode Statistics between 1991 and 1995 [15]. The observed mortality is for operations performed in the Bristol Royal Infirmary, whereas the expected mortality rates are the median rates in 11 other centres; no adequate risk-adjustment procedure was available [16].

Shipman

We consider two analyses of mortality rates for men and women aged 65 years and over in the practice list of Dr Harold Shipman using data provided by Baker [17]: firstly, death between 1977 and 1998 taking place in the patient's home or in practice premises and having certificates signed by the GP himself, as compared with the rate of such certificates signed by a sample of local GPs, and secondly, all deaths in Shipman's practice between 1987 and 1998, as compared with age-adjusted expected rates for England and Wales.

Cardiac surgeons

There are many available data sets of mortality rates for adult cardiac surgery [18]. We took a set analysed previously using the risk-adjusted CUSUM technique [10]. These comprise the 30-day mortality from individual coronary artery bypass graft operations carried out between 1994 until 1998, risk-adjusted using the Parsonnet system which has been re-calibrated on surgery in the same unit between 1992 and 1993 [10]. Data for two surgeons are provided.

Results

Bristol

The cumulative excess mortality in Bristol from the two data sources is displayed in Figure 1A, which is transformed in Figure 1B to the SPRT plot for monitoring. There are 12 centres in England performing such surgery, so if such a procedure were to have been used for prospective monitoring it may have been reasonable to demand α's of around 0.01 for 'alert' and 0.001 for 'alarm'. The Cardiac Surgical Register data crossed the $\alpha = 0.05$ boundary in 1989, and the 0.001 boundary in 1991, whereas the Hospital Episode Statistics results crossed the 0.01 boundary in 1993 and passed 0.001 in 1994.

Shipman

The cumulative excess of deaths certified by Shipman at home or in practice premises from 1977 is shown in Figure 2A, revealing a steadily increasing pattern for females, particularly accelerating after 1995. The final totals closely mirror the conclusion of the public inquiry [2], that Shipman killed, or probably killed, 180 women and 55 men aged 65 years or over. There are around 9000 general practices involving 27 000 doctors in England, suggesting that any monitoring procedure would want to use stringent criteria such as $\alpha = \beta = 0.000001$. Figure 2B shows that the graph for females dying at home or in the practice crossed this limit in 1985, at which point there were 41 excess female deaths. At the point where Shipman was arrested, in 1998, this excess had grown to 174.

However, compared with England and Wales, mortality rates irrespective of place of death, excess total mortality on Shipman's list was broadly in line with national rates between 1988 and 1994 and only started accumulating in 1995 for women (Figure 3A). Figure 3B shows that the cumulative LLR thus builds up substantial credit and therefore it is not until 1997 that an 'alarm' threshold of 0.000001 is crossed. However, by 1992 the test would have confidently concluded that Shipman did not have a raised mortality rate and the monitoring procedure could therefore have been restarted, and in Figure 3C we see this could have led to an alarm being raised at the end of 1996, when there were 67 excess deaths in females aged over 65 years, compared with 119 by 1998.

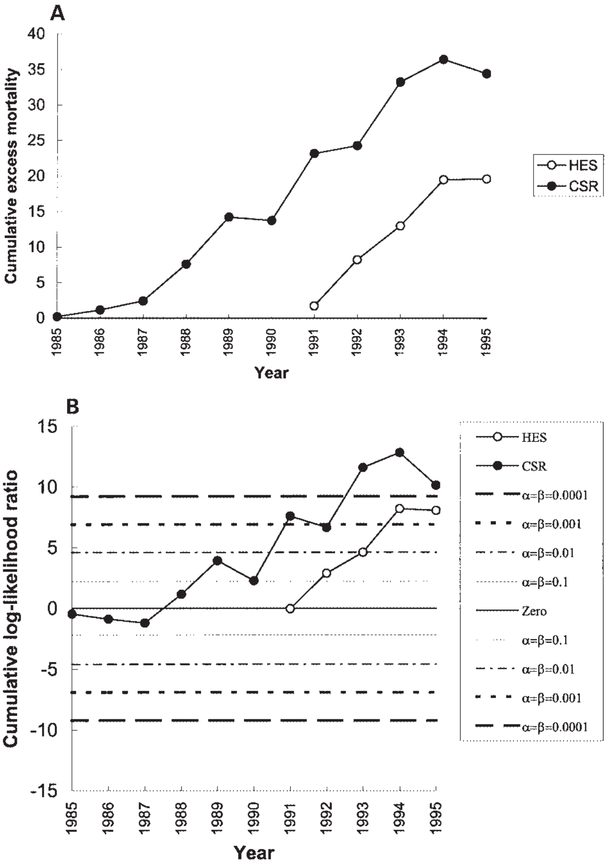

Figure 1 (**A**) Cumulative excess mortality in Bristol for cardiac surgery on children under 1 year of age. (**B**) Sequential probability ratio test for detection of a doubling in the odds on mortality. HES, Hospital Episode Statistics; CSR, Cardiac Surgical Register; α, false positive error rate; β, false negative error rate.

Cardiac surgeons

In the data sets selected for illustration, surgeon 2 has some excess mortality early on in the series but made up for this later on (Figure 4A), whereas surgeon 1 showed some small improvement in mortality over that expected. The SPRT plot in Figure 4B shows no evidence for increased risk and an eventual confident rejection of that hypothesis for both surgeons. Figure 4C, similarly, shows no evidence for substantially decreased risk: in fact there is a confident rejection of a decrease in risk at around case 100 for surgeon 2, at which point it might be reasonable to restart the monitoring procedure.

Discussion

The techniques discussed in this paper can be seen as a natural development from control charts and CUSUMs, and

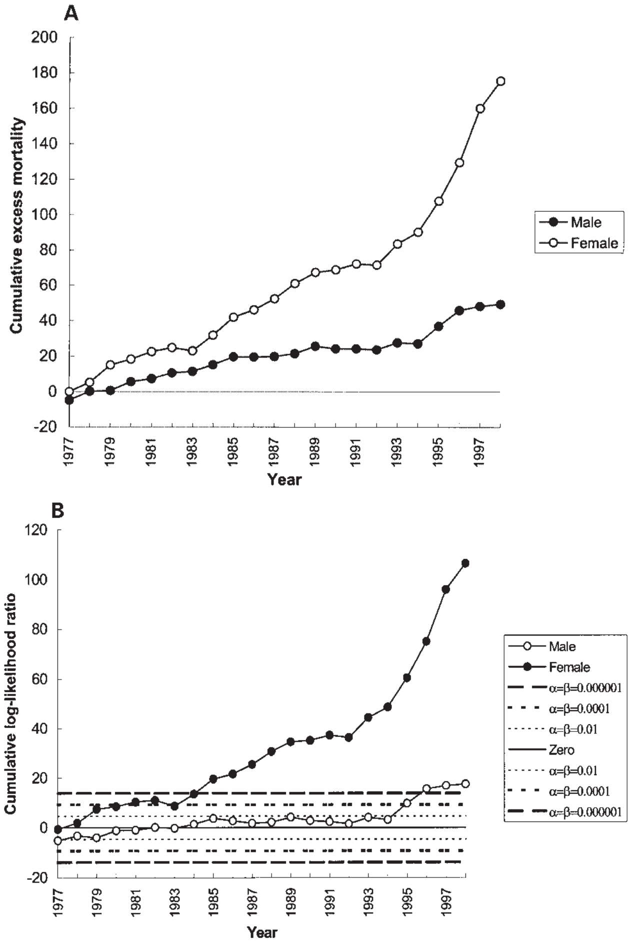

Figure 2 (**A**) Cumulative excess death certificates signed by Dr Shipman: age >64 years and death in home/practice. (**B**) Sequential probability ratio test for detection of a doubling in mortality risk: age >64 years and death in home/practice. α, false positive error rate; β, false negative error rate.

naturally complement intuitively attractive plots of cumulative observed–expected mortality [5,6]. They are easily implemented in spreadsheet programs.

Bristol

The analysis suggests that the divergent performance for Bristol might have been detected earlier using a simple statistical technique applied to routinely collected data. However, we note that in practice the Cardiac Surgical Register results were not available to centres until at least a year has elapsed, whereas Hospital Episode Statistics were not used for monitoring purposes.

Shipman

Our first analysis may well over-estimate the mortality attributable to wrongdoing, as there are many innocuous reasons why a GP might sign more certificates for deaths occurring at home, say because of his care policy for terminally ill patients. In addition, there were only between four and six control GPs in any year, so the control group is local but limited and possibly non-representative. The second comparison uses national death rates as controls, but could be insensitive if Shipman's victims tended to be in poor health and hence his actions did not increase overall death rates. This appears to have been the pattern until around 1995. It is important to note that the data used in our analysis were specially put together after Shipman's conviction and no such routine monitoring was in place: in particular, death certificates do not carry details of the GP and linkage must

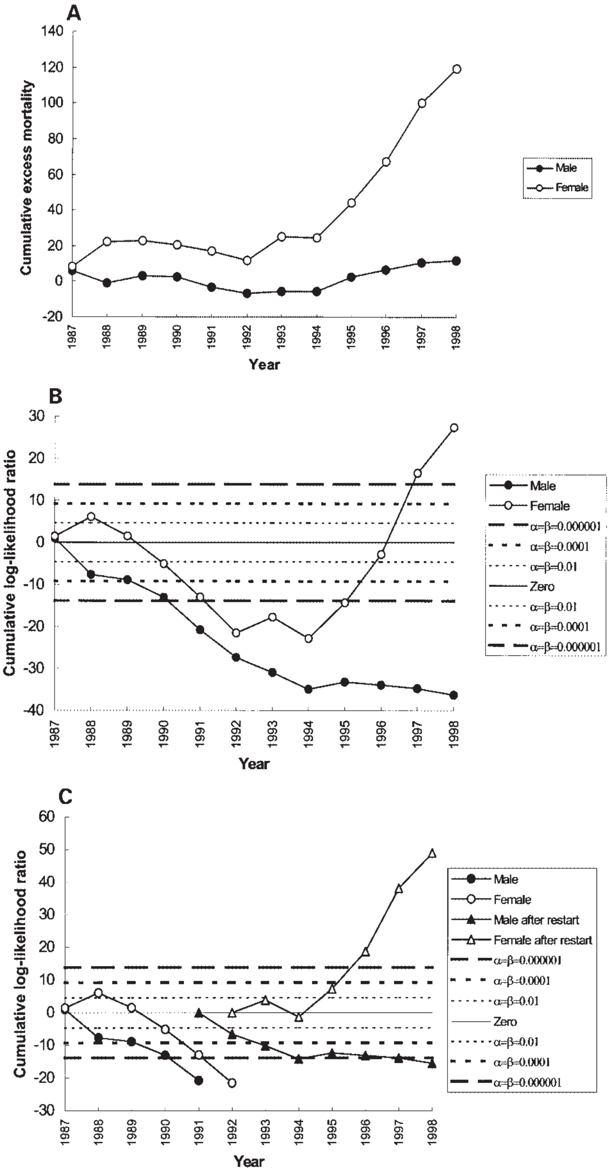

Figure 3 (**A**) Cumulative excess mortality for age >64 years in Dr Shipman's practice, compared with England and Wales. (**B**) Sequential probability ratio test for detection of a doubling in mortality risk: age >64 years, compared with England and Wales. (**C**) Sequential probability ratio test for detection of a doubling in mortality risk allowing for restarts: age >64 years. α, false positive error rate; β, false negative error rate.

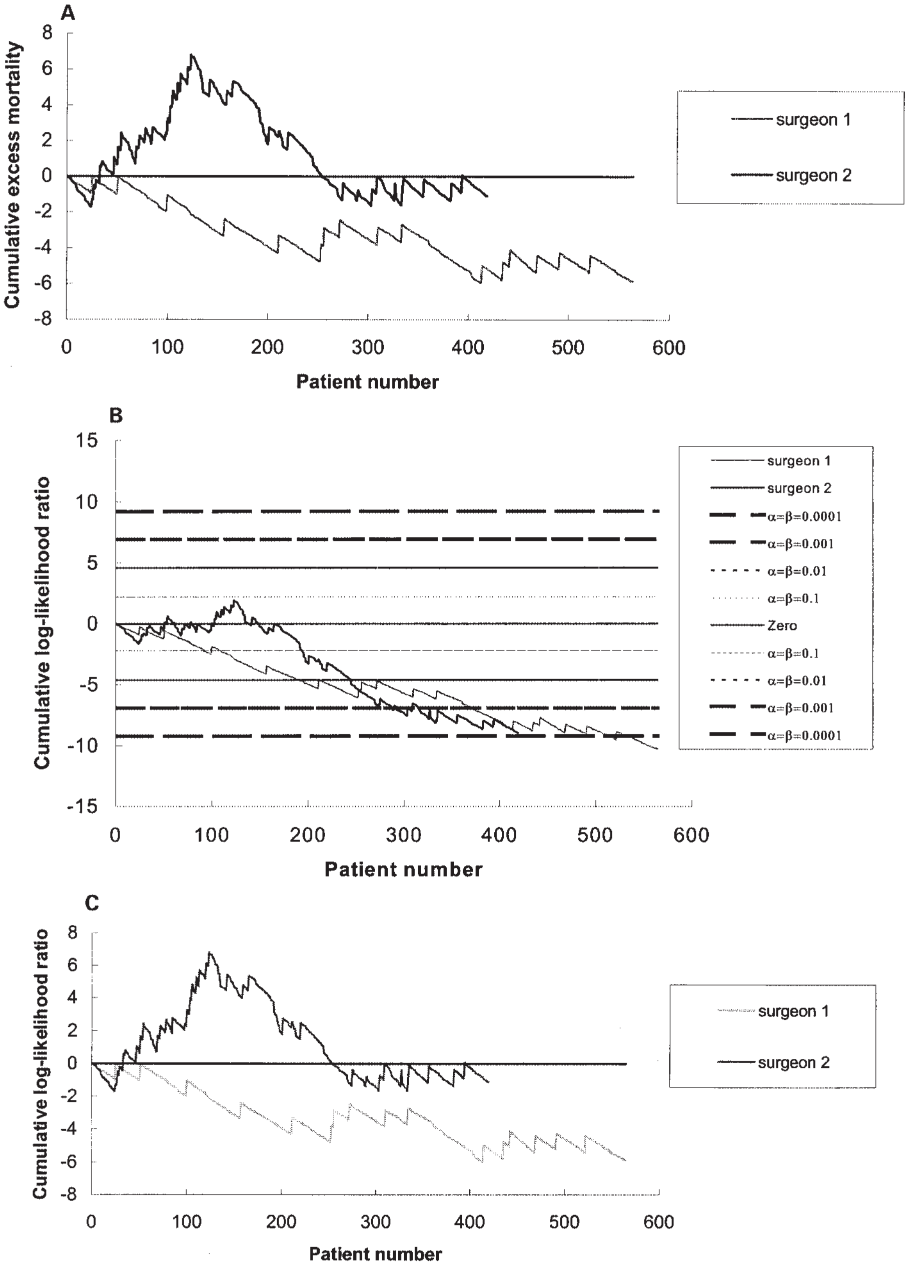

Figure 4 (**A**) Sequential probability ratio test for detection of a doubling in the odds on mortality, assuming constant baseline risk. (**B**) Sequential probability ratio test for detection of a doubling in the odds on mortality. (**C**) Sequential probability ratio test for detection of a halving in the odds on mortality. α, false positive error rate; β, false negative error rate.

be made through the NHS Central Register. Nevertheless, the results suggest that a simple monitoring procedure could have led to earlier detection of Shipman's crimes and a substantial saving of lives.

Cardiac surgeons

This example demonstrates the importance of adjustment for pre-operative risk in surgical monitoring. The SPRT that assumes constant baseline risk, that is, that does not adjust for risk, suggests that the surgeons are performing divergently. However, the risk-adjusted version of the SPRT, which does take into account factors that are beyond a surgeon's control, indicates that the surgeons are in fact performing similarly.

The issue of 'restarts' is quite complex. Crossing the lower boundary confirming 'performance as expected' is reassuring but otherwise is not of great interest, and it seems natural to begin a new monitoring session at that stage in order to retain sensitivity to changes in performance. However, a series of such restarts increases the overall chance of a Type I error eventually to 1 and, because in effect we are avoiding 'accepting' the null hypothesis, decreases the Type II error eventually to 0. If we further encourage restarts by making β larger, then the behaviour of the SPRT begins to closely resemble that of the risk-adjusted CUSUM [10,19]. This latter technique uses the same form of LLR as the SPRT for monitoring, but has no concept of accepting the null hypothesis: the scheme is restarted each time it crosses 0, and the boundaries are set in a pragmatic way. Both SPRTs and risk-adjusted CUSUMs warrant further research, particularly as the approximation underlying the boundaries of the SPRT may be questionable for Poisson data.

Implementation of such monitoring systems requires good-quality data, as well as specification of thresholds and expected performance. If a SPRT is being used within an institution then setting thresholds requires consideration of the seriousness of each type of error. However, if a central audit body is examining multiple SPRTs from a number of institutions then they need to additionally take into account the multiple comparisons being made. As a rule-of-thumb we have suggested the simple Bonferroni procedure of dividing the error rates by the number of institutions being monitored, but this requires further investigation.

The risk-adjustment scheme used to derive expected performance can cause difficulties. For example, if monitoring GPs, should comparison only be made with other practitioners in the immediate neighbourhood, or those with similar socio-economic practice lists? An appropriate baseline is essential for fair comparisons. Finally, Frankel and colleagues claim that a monitoring system may only detect really extreme divergence [20]. We have somewhat arbitrarily identified a doubling of risk as being 'important', and shown sensitivity of the procedure even allowing for multiple comparisons.

The SPRT is one of the simplest prospective monitoring schemes: slightly more sophisticated developments might include the risk-adjusted CUSUM [10], systematic down-weighting of historical cases or shrinkage estimation of rates towards the average [4,21,22]. Such formal statistical methodology may aid routine monitoring of clinical performance.

Acknowledgements

We are grateful to Professor Vern Farewell for many useful discussions.

References

1. BRI Inquiry Panel. *Learning from Bristol: The Report of the Public Inquiry into Children's Heart Surgery at the Bristol Royal Infirmary 1984–1995*. London, UK: The Stationery Office, 2001. Available from http://www.bristol-inquiry.org.uk/final_report/.
2. *Shipman Inquiry: The First Report*. London, UK: The Stationary Office, 2002. Available from http://www.the-shipman-inquiry.org.uk/reports.asp.
3. Mohammed M, Cheng K, Rouse A, Marshall T. Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons. *Lancet* 2001; **357:** 463–467.
4. 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:** 914–924.
5. 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:** 1128–1130.
6. Poloniecki J, Valencia O, Littlejohns P. Cumulative risk-adjusted mortality chart for detecting changes in death rate: observational study of heart surgery. *Br Med J* 1998; **316:** 1697–1700.
7. Lawrance R, Dorsch M, Sapsford R *et al.* Use of cumulative mortality data in patients with acute myocardial infarction for early detection of variation in clinical practice: observational study. *Br Med J* 2000; **323:** 324–327.
8. 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.
9. McPherson CK. Statistics: the problem of examining accumulating data more than once. *N Engl J Med* 1974; **290:** 501–502.
10. Steiner SH, Cook RJ, Farewell VT, Treasure T. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; **1:** 441–452.
11. Barnard GA. Sequential tests in industrial statistics (with discussion). *J R Statist Soc* 1946; **8(suppl.)**: 1–26.
12. Wald A. Sequential tests of statistical hypotheses. *Ann Math Statist* 1945; **6**: 117–186.
13. Armitage P. Sequential tests in prophylactic and therapeutic trials. *Q J Med* 1954; **23:** 255–274.
14. Bartholomay AF. The sequential probability ratio test applied to the design of clinical experiments. *N Engl J Med* 1957; **256:** 498–505.
15. Spiegelhalter DJ, Evans S, Aylin P, Murray GD. Overview of statistical evidence presented to the Bristol Royal Infirmary Inquiry concerning the nature and outcomes of paediatric cardiac surgical services at Bristol relative to other specialist centres from 1984 to 1995. Bristol Royal Infirmary Inquiry 2000. Available from http://www.bristol-inquiry.org.uk/final_report/annex_b/images/Spiegelhalteretal_O_statev1.pdf (Accessed August 10 2001).
16. Aylin P, Alves B, Best N *et al.* Comparison of UK paediatric cardiac surgical performance by analysis of routinely collected data 1q984-96: was Bristol and outlier? *Lancet* 2001; **358:** 181–187.
17. Baker R. Harold Shipman's Clinical Practice 1974–1998: A Review Commissioned by the Chief Medical Officer. London: The Stationery Office, 2001.
18. Keogh BE, Kinsman R. National Adult Cardiac Surgical Database Report 1999–2000. The Society of Cardiothoracic Surgeons of Great Britain and Ireland.
19. Grigg O. *A Comparison of Methods for Monitoring Surgical Performance*. Master's thesis, Department of Statistical Science, University College London, 2001.
20. Frankel S, Sterne J, Smith G. Mortality variations as a measure

of general practitioner performance: implications of the Shipman case. *Br Med J* 2000; **320:** 489.

21. Goldstein H, Spiegelhalter DJ. Statistical aspects of institutional performance: league tables and their limitations (with discussion). *J R Stat Soc Series A* 1996; **159:** 385–444.

22. Christiansen C, Morris C. Improving the statistical approach to health care provider profiling. *Ann Intern Med* 1997; **127:** 764–768.

Accepted for publication 28 August 2002

Appendix

Here we describe calculation of the LLR weights for charts detecting a doubling in the odds or risk ratio (such as the charts illustrated in this paper), with reference to the cardiac surgery and Shipman examples. Also, a table giving the values of the boundaries *a* and *b* for various equal α and β is provided, to illustrate the use of the Wald equations given in the Materials and methods section of this paper.

Suppose we observe data $x_1, x_2, \ldots, x_n$ and wish to compare two hypotheses H_0, H_1. Then the most powerful test is based on the 'log-likelihood ratio' LLR $= \Sigma_i \ln(\mathrm{P}_1(x_i)/\mathrm{P}_0(x_i))$ where $\mathrm{P}_0(x)$ is the probability of observing the data under H_0 and so on.

Consider patient by patient monitoring, such as in the cardiac surgery example (Bernoulli data). Given that the preoperative risk is p, a doubling in the odds on death is optimally detected by adding the score $-\ln(1+p)$ to the running score total if the patient survives, and $0.69 - \ln(1+p)$ if the patient dies. The value 0.69 is equal to ln(2). For detecting a halving of the odds on death, this value would be replaced by ln(1/2). To adjust for risk, we can allow the value of p to depend on individual patient risk factors such as age, sex and diabetes status.

For monitoring of aggregate data, such as the Shipman data (Poisson data), a doubling in risk of death is optimally detected by adding $0.69O - E$ to the running LLR, where O is the observed number of deaths and E the expected number of deaths over a fixed time interval. Note that this is a simple adjustment to the cumulative 'excess mortality' plot of $O-E$. To adjust for risk, we simply use expected counts based on a suitable comparator (England and Wales averages, for example) for the risk categories that we are interested in.

Table 1 Thresholds for the running log-likelihood ratio for different values of α and β

α[1]	β[2]	Lower threshold, *a*	Upper threshold, *b*
0.05	0.05	−2.94	2.94
0.01	0.01	−4.60	4.60
0.005	0.005	−5.29	5.29
0.001	0.001	−6.91	6.91
0.0001	0.0001	−9.21	9.21
0.00001	0.00001	−11.51	11.51
0.000001	0.000001	−13.82	13.82

Note: H_1 is rejected when log-likelihood ratio is less than *a* and H_0 is rejected when log-likelihood ratio is greater than *b*.

[1] False positive (Type I) error rate.

[2] False negative (Type II) error rate.