Monitoring surgical and medical outcomes: the Bernoulli cumulative SUM chart. A novel application to assess clinical interventions

G Leandro, N Rolando, G Gallus, K Rolles, A K Burroughs

Background: Monitoring clinical interventions is an increasing requirement in current clinical practice. The standard CUSUM (cumulative sum) charts are used for this purpose. However, they are difficult to use in terms of identifying the point at which outcomes begin to outside recommended limits.

Objective: To assess the Bernoulli CUSUM chart that permits not only a 100% inspection rate, but also the setting of average expected outcomes, maximum deviations from these, and false positive rates for the alarm signal to trigger.

Methods: As a working example this study used 674 consecutive first liver transplant recipients. The expected one year mortality set at 24% from the European Liver Transplant Registry average. A standard CUSUM was compared with Bernoulli CUSUM: the control value mortality was therefore 24%, maximum accepted mortality 30%, and average number of observations to signal was 500—that is, likelihood of false positive alarm was 1:500.

Results: The standard CUSUM showed an initial descending curve (nadir at patient 215) then progressively ascended indicating better performance. The Bernoulli CUSUM gave three alarm signals initially, with easily recognised breaks in the curve. There were no alarms signals after patient 143 indicating satisfactory performance within the criteria set.

Conclusions: The Bernoulli CUSUM is more easily interpretable graphically and is more suitable for monitoring outcomes than the standard CUSUM chart. It only requires three parameters to be set to monitor any clinical intervention: the average expected outcome, the maximum deviation from this, and the rate of false positive alarm triggers.

S

ystems to monitor clinical performance have gained increasing importance in medical practice. In the United Kingdom, since the Bristol Royal Infirmary Inquiry there has been a new impetus to monitor outcomes of surgical and medical interventions. Assessing medical or surgical practice can take advantage of techniques studied for controlling a manufacturing process in industry to create a product. Maintaining an optimum is necessary. Medical audit aims to monitor the variation around a predetermined optimum and ideally alarm signals should be set at the point when performance goes beyond acceptable standards. However, most audit procedures are retrospective analyses of outcome such as for mortality or important morbidity. This differs from manufacturing industry where real time assessment of the production process occurs.

Any process, whether in a biomedical or non-biomedical field can be subject to inherent deviations from the optimum or from designated limits. These deviations may lead to defective end products or in the medical field defective patient care. Monitoring, which is a process able to identify deviation, and then being able to act if this deviation exceeds certain limits, plays an important part in avoiding adverse consequences and maintaining optimal performance. In production processes the techniques most often used to monitor and control are referred to as statistical process control (SPC), which today is fully accepted as a key element in modern manufacturing industry.

In biomedical science, statistical evaluation usually considers only the first or the last point of a dataset. These statistical tests do not take into account what has happened in the past—that is, data between the first and the last point of a dataset are not considered. In contrast the SPC permits use of information at more than one time point—that is, sequential testing. SPC has been used in biological applications for surveillance of rare events, but only rarely has it been used in biomedicine, an example being the monitoring of congenital defects by one of us (GG). This is mainly because it has proved difficult to bring together the underlying theories with appropriate procedures of calculation, as these are somewhat cumbersome and not easily applied to clinical processes. Among the sequential tests, the cumulative sum (CUSUM) test has had the widest use, in both surveillance and quality control, and in evaluating sequential measures or changes over time. Williams et al used the simple CUSUM to illustrate its use in monitoring success with colonoscopy and endoscopic retrograde cholangiopancreatography, and its application in monitoring individual clinicians. The method assumes a fixed and defined limit of probability of success of the endoscopic procedure. The authors felt that this simple CUSUM technique should be used as a means of monitoring performance, both for training in medical or surgical techniques and in monitoring to ensure the maintenance of standards.

Lovegrove et al evaluated the probability of death after surgical cardiothoracic procedures, which varied for every patient added cumulatively to a series, according to various risk factors such as comorbidity and surgical expertise.

Both these studies have cumulative series, detailing success or failure. However, an evaluation of the process over time is in practice made by a visual examination of the graphical representation. There are boundary lines traced on

Abbreviations: CUSUM, cumulative sum; SPRT, sequential probability ratio test; ELTR, European Liver Transplant Registry
the CUSUM graph, which are placed to define when the slope changes beyond acceptable limits, but the methodology used is cumbersome and is not sufficiently precise, thus giving the probability of too many false alarms.

What is missing is a formal and robust statistical test to assess medical or surgical performance over time and to have a monitoring system that gives very few false alarms. While further refinement has led to evaluating risk adjusted outcomes, particularly for mortality, these as yet have not entered clinical practice. This is partly because not all datasets contain the relevant risk factors, or datasets are felt to be unreliable in terms of recording risk factors and partly because a consensus needs to be reached on the validity of the risk factors for example in cardiothoracic surgery,

Monitoring methodology

Recently monitoring protocols of biomedical processes have benefited from developments in computation of statistical procedures, which are specifically designed to evaluate alarm signals. These result in a highly accurate and comparable methodology in the form of the Bernoulli CUSUM control chart, as a type of statistical process control, to a clinical process for which outcomes are known—that is, the number with suboptimal outcome—and for which the limits for an unacceptable deviation from the baseline can be defined, so that alarm signals regarding unacceptable deviation beyond the specified limits can be detected.

METHODS

The details of the theory, terminology, and calculations are shown in the appendix. As an example of a clinical intervention, we chose a consecutive cohort of adult patients undergoing their first liver transplant at a single centre, with death at one year as the primary outcome measure. The reasons for this choice are the following: (1) high intrinsic mortality rate and thus a high event rate, compared with other surgical procedures; (2) a comparatively standardised procedure within a single centre with less surgical variability; (3) the existence of a large international published registry with validated data, the European Liver Transplant Registry (ELTR), which has been published with one year mortality figures for patients with and without associated risk factors for mortality. The total population was 19370 adults; (4) similarities between the reference registry cohort and the single centre cohort at the Royal Free Hospital. The database covered a similar time period and there were similar proportions of the following characteristics respectively: orthotopic liver transplants (99.7% v 99.9%), fulminant liver failure (11% v 10%), cancer (10% v 11%), re-transplantation (12% v 10.5%) the last three being risk factors for one year mortality.

RESULTS

The consecutive series of first liver transplants at the Royal Free Hospital between 1 October 1988 (the start of the programme) and 31 August 2002 were 674. The median age was 49 years and 60% were male. There were 71 re-transplants (eight of these had a third transplant) and 12 had a combined liver and kidney transplant.

The expected one year mortality was considered to be 24% for the whole population as this was the actual mortality for adults in the ELTR in the report published in 2000. Although our cohort extended to 2002, and mortality after
liver transplantation is improving,11 the average 24% mortality is used to give a working example with a real dataset. Therefore, following the suggestion of Williams et al15 for every patient alive at one year the probability of death \( (p_0 = 0.24) \) was added to the CUSUM, and for every death within one year, the probability of being alive \( (1 - p_0 = 0.76) \) was subtracted from the CUSUM. Figure 1 shows the cumulative course of the Royal Free cohort according to the above calculations. A declining course can be seen in the first part and an ascending course in the second.

The second chart (fig 2) displays the same dataset using the Bernoulli CUSUM control chart. The parameters used are: probability of death = \( p_0 = 0.24 \) (in control value of \( p \)). The maximum value of \( p \), that is the maximum permitted mortality at one year, beyond which the process is defined as being outside of the specified parameters was considered to be 30%—that is, \( p_1 = 0.30 \) (out of control value of \( p \)). The average number of observations to signal, ANOS \( (p_0) \) was set to the value 500. The ANOS \( (p_0) \) is the number of observations that would contain one false alarm—that is, a false positive—and is explicitly stated a priori. The alarm signal means that the process is out of control, in other words that there is a factor or several factors that are leading to an unacceptable deviation from the specified limits. In manufacturing industry, the production process would be stopped and machines would be re-gauged. In medical audit this would be the time to evaluate all potential factors that could lead to an out of control result and to proceed accordingly.

In figure 2 three alarm signals (given when the sum reaches the value 24) are represented by interruptions in the line in the first 138 transplants. In our series the out of control process was probably due to several factors: case mix, surgical technique, case volume, and learning curve in integrating the multidisciplinary teams involved in this complex surgery and postoperative care.21 Clearly in the early years of the programme the expected average mortality at one year would have been lower than the average 24%, and in the later years higher. However, what the earlier or later expected average mortalities are, cannot be derived for the earlier period from the ELTR database20 nor estimated for the later period. Nevertheless the fact that alarm signals are only seen in the early period and then results progressively improve to above the average suggests improving performance in common with many transplant centres.19

When assessing the patients from the Royal Free cohort without risk factors (ignoring centre size effect based on ELTR data that give a risk ratio of 1.46 compared with baseline mortality), which were 42, and evaluating one year mortality with \( p_0 = 0.15 \) derived from the ELTR dataset, and considering \( p_1 \) as equal to 0.20 and ANOS \( (p_0) = 500 \), then...
There are no alarm signals at any time (fig 3). This suggests that the initial alarm signals in the whole cohort were less likely to be attributable to surgical techniques or initial perioperative care as part of the learning curve, as one might expect a similar trend in patients without known risk factors. Recipient selection or use of marginal donors may have been reasons for the poorer results and thus the alarm signals. In addition as the average probability of death would have been higher in the early years—1988–1992—it would seem that increasingly improved results with time in our centre and in others (19) may be attributable to better management of worse risk patients with more risk during and after liver transplantation—an effect of increasing experience.

The difference between figure 1, using the CUSUM graphical representation, which gives only a visual assessment, and the Bernoulli CUSUM approach of figure 2, is that the graphical representation in figure 1 suggests the nadir was reached at number 215 and then there was a continuous improvement, whereas in figure 2 it can be seen that no alarm signals occurred after transplant 148—that is, the process had been under control from an earlier time point.

Moreover in the phase of the descending slope using the simple CUSUM representation, there are no fixed points where one can decide to audit the problem to establish the causes of the deviation from the given parameters.

**DISCUSSION**

Medical and surgical interventions need a monitoring system, which can signal when the process is getting out of control, as is the case in non-medical fields. How to devise such monitoring is a matter of current and topical interest, particularly in the UK, given the results of the Bristol Royal Infirmary Inquiry, which has resulted in applying graphical CUSUM control charts and derived risk adjusted models (variable life adjusted display) in cardiothoracic procedures, although there are difficulties in setting thresholds. There also have been questions regarding the reliability of the risk scores as well as comparisons of crude and risk adjusted mortality. Currently crude mortality rates are still being evaluated. The quality and credibility of the UK cardiac surgery database has been assessed as needing improvement making accurate risk adjusted mortality more difficult to assess. For other surgical procedures mortality control charts may not be the best way to assess or compare performance. Thus hospital mortality charts for common surgical procedures (and medical conditions) have been considered crude and at worst misleading reflections of quality of care. Despite these problems cardiac surgeons in the UK will be assessed according to bypass surgery success, a decision that is a by product of the momentum generated to demonstrate quality control “in action”. Thus the improvement in statistical computation that has occurred recently as well as improved graphical display are useful, even if solely applied to evaluations of unadjusted mortality.

However, the publications using descriptive CUSUM charts, and derived variable life adjusted display (VLAD) charts have the disadvantage of not displaying the precise point of the alarm signal, and often an assessment can only be made by eye-balling the shape of the curves. Only recently has statistical methodology been applied to monitor medical or surgical outcomes when risk adjustment cannot be manipulated adequately or to evaluate optimal performance with pre-set thresholds.

In our study we have evaluated a new monitoring system, the Bernoulli CUSUM chart, which has a 100% inspection rate, using as an example of a clinical intervention, a surgical procedure—liver transplantation. We considered mortality at one year as an outcome in a consecutive series of adult patients from the start of the programme in 1988 to 2002.

This monitoring system requires the setting of three values: the average probability of death after transplantation \( p_0 \), the probability of the failure rate considered unacceptable in current practice \( p_1 \), and the probability of a false alarm signal during monitoring \( ANOS(p_0) \) at 1:500. No alarm signals have occurred.
display, with the interrupted lines in the Bernoulli CUSUM chart, representing that the alarm signals have been reached. The first alarm signal occurs before the nadir in the standard CUSUM chart. In addition the chart shows an earlier return to a system under control.

Secondly, the statistical methodology behind the Bernoulli CUSUM control chart is validated and accepted in fields outside medicine and indeed is the most applicable to interventions in medicine as it uses a 100% inspection rate. We believe that this monitoring system, as it permits simple setting of thresholds for performance and error in the alarm signals, overcomes many of the disadvantages of current systems and is a step forward in assessing outcome of surgical and medical procedures.

In addition the use of this methodology has several potential applications over and above monitoring a procedure to establish if the outcomes are maintained within set thresholds based on average performance published in the literature (as in our example). One application could be to delineate periods of worse performance (even before alarm signals are reached) permitting a detailed evaluation of a particular series of patients, to try to understand factors that may have led to this. A second application could be that if the monitoring shows a prolonged period of stability or improvement, the thresholds ($p_0$ and $p_1$) could be modified to evaluate what performance level has been reached for the particular procedure under scrutiny. In our example, evaluating the more recent, second half of the cohort (fig 4), the alarm signals are not reached until thresholds are set at an average one year mortality of 16% ($P_0$) and maximum alarm signals are not reached until thresholds are set at an average one year mortality of 16% ($P_0$) and maximum.

To calculate the value of $\gamma_b$, it is necessary to specify a value ($p_0$) which represent the out-of-control value of $p$, $p_1$, is also the value of $p$ which should be detected as soon as possible. Given the in-control value $p_0$ and the out-of-control value $p_1$, we can obtain the constants $r_1$ and $r_2$ by

$$r_1 = -\log \left( \frac{1 - p_1}{1 - p_0} \right)$$
$$r_2 = \log \left( \frac{p_1 (1 - p_0)}{p_0 (1 - p_1)} \right)$$

and then the value of $\gamma b$ as

$$\gamma_b = \frac{r_1}{r_2} = \frac{1}{m}$$

where $m$ is an integer.

The average number of observations to signal (ANOS) is obtained by the so called corrected diffusion approximation and is:

$$ANOS(p_0) \approx \frac{\exp [h_{20}^{*} r_2] - h_{20}^{*} r_2 - 1}{|r_2 p_0 - r_1|}$$

where $h_{20}^{*}$ is an adjusted value of $h_B$ and was obtained by the formula

$$h_{20}^{*} = h_B + \epsilon(p_0) \sqrt{p_0(1-p_0)}$$

where $q_0 = 1 - p_0$ and

$$\epsilon(p) \approx 0.41 - 0.0842 (\log(p))^3 - 0.00376 (\log(p))^4 - 0.000008 (\log(p))^5$$

if $0.01 \leq p \leq 0.5$ or

$$\epsilon(p) \approx \frac{1}{5} \left( \sqrt{\frac{1 - p}{p}} - \sqrt{\frac{p}{1 - p}} \right) \text{ if } 0 < p < 0.01$$

The ANOS ($p_0$) is the number of observations that would contain one false alarm and is explicitly stated a priori—that is, a false positive.

To know how fast a shift from $p_0$ to $p_1$ will detected, the CD approximation to the ANOS when $p = p_1$ is

$$ANOS(p_1) \approx \frac{\exp [h_{20}^{*} r_2] + h_{20}^{*} r_2 - 1}{|r_2 p_1 - r_1|}$$

REFERENCES
27 Dyer O. Heart surgeons are to be rated according to bypass surgery success. BMJ 2003;326:1053.