Monitoring the evolutionary process of quality: Risk-adjusted charting to track outcomes in intensive care

David A. Cook, MBBS; Stefan H. Steiner, PhD; Richard J. Cook, PhD; Vern T. Farewell, PhD; Anthony P. Morton, MD

Objective: **To present graphical procedures for prospectively monitoring outcomes in the intensive care unit.**

Design: **Observational study: risk-adjusted control chart analysis of a case series.**

Setting: **Tertiary referral adult intensive care unit: Princess Alexandra Hospital, Brisbane, Australia.**

Patients: **A total of 3398 intensive care unit admissions from January 1, 1995, to January 1, 1998.**

Conclusions: **Risk-adjusted process control charting procedures for continuous monitoring of intensive care unit outcomes are proposed as quality management tools. A modified Shewhart *p* chart and cumulative sum process control chart, using the Acute Physiology and Chronic Health Evaluation III model mortality prediction for risk adjustment, are presented. The risk-adjusted *p* chart summarizes performance at arbitrary intervals and plots observed against predicted mortality rate to detect large changes in risk-adjusted mortality. The risk-adjusted cumulative sum procedure is a likelihood-based scoring method that adjusts for estimated risk of death, accumulating evidence from outcomes of all previous patients. It formally tests the hypothesis of a change in the odds of death. In this application, we detected a decrease from above to predicted risk-adjusted mortality. This was temporally related to increased senior staffing levels and enhanced ongoing multidisciplinary review of practice, quality improvement, and educational activities. Formulas and analyses are provided as appendices. (Crit Care Med 2003; 31:1676–1682)**

KEY WORDS: risk adjustment; cumulative sum procedure; Shewhart *p* chart; control chart; monitoring; outcome; performance; intensive care

Outcome measurement is a cornerstone of quality improvement. Continuous outcome monitoring provides an approach to guide improvements in quality of care by providing feedback about the overall effects of changes in practice. Monitoring can be used to detect a change in outcomes so the process of care can be examined, reinforcing beneficial practices and eliminating factors that degrade performance. Alternatively, monitoring can be used to prospectively evaluate single or multiple interventions. In a culture of change, driven by innovation and the incorporation of empirical research-based evidence (evidence-based medicine), continuous monitoring of outcomes has the potential to direct the evolution of practice toward higher quality.

In medical applications, as in industry, process performance may be better understood with evaluation of the consequences of process interventions and provision of sequential information (1). Some quality surveillance applications have been described in the medical literature, including monitoring clinical outcomes in colonoscopy (2–4), pediatric cardiac surgery (5), and adult thoracic surgery (6). Continuous surveillance accumulates evidence to detect subtle changes, may detect cyclic changes, and can offer contemporaneous feedback and analysis.

Process monitoring may include Shewhart *p* charts and the cumulative sum (CUSUM) procedure. The *p* chart can detect large changes in a rate of occurrence, and the CUSUM is suited to detecting small persistent changes in an event rate over time (7).

Manufacturing and industrial applications typically have homogeneous input specifications, so traditional control chart approaches make no adjustment for different risks of failure. In contrast, patients referred to intensive care units (ICU) are heterogeneous in their clinical presentation and physiology. *Case-mix* is a term that broadly describes the considerable variability in patients' condition and severity. Incorporation of a validated risk of death adjustment to control chart analysis is conceptually similar to modeling the effects of confounding variables in a controlled study. In risk-adjusted (RA) outcome monitoring, the confounding factor is mortality risk.

In the RA *p* chart, the observed mortality rate is compared with control limits estimated around the predicted mortality rate accounting for case-mix, patient numbers, and random variation (Appendix 1). As with outcome analysis with the standardized mortality ratio, repeated sample periods will increase the probability of false alarm in a predictable manner.

Steiner et al. (8, 9) developed a RA CUSUM procedure for continuous monitoring based on the likelihood ratio. They propose an efficient graphical method for identifying when there is either a substantial decrease or increase in RA mortality (Appendixes 2 and 3).

In this report, an RA *p* chart and a two-sided RA CUSUM chart are applied to

From the Intensive Care Unit (DAC) and the Infection Management Services (APM), Princess Alexandra Hospital, Brisbane, Australia; the Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario, Canada (SHS, RJC); MRC Biostatistics Unit, Cambridge, UK (VTF).

Supported, in part, by the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research.

DOI: 10.1097/01.CCM.0000065273.63224.A8

data from an ICU data set to monitor ICU outcomes. The RA is based on the locally validated Acute Physiology and Chronic Health Evaluation (APACHE) III (10) model. This is the first report of outcome surveillance procedures proposed to continuously monitor local changes in RA mortality performance in ICU.

MATERIALS AND METHODS

The data set is drawn from 3398 consecutive ICU admissions (3159 hospital admissions from January 1, 1995, to January 1, 1998) to the Princess Alexandra Hospital in Brisbane, Australia. The Princess Alexandra Hospital ICU provides medical and surgical critical care services to an 858-bed adult metropolitan teaching hospital, which is the regional center for trauma, major surgery, medical subspecialties, and psychiatry.

The ICU mortality analysis included all eligible admissions to the ICU, including readmissions. The hospital mortality analysis excluded all ICU readmissions during an episode of hospitalization to prevent double counting of outcomes. For each admission, estimates of in-ICU and in-hospital mortality rates were calculated with the APACHE III equation (10). The details of local performance and model validation have been described (11). The receiver operating characteristic curve area was 0.90, and calibration curves and goodness-of-fit analysis indicated that the models with proprietary adjustments for hospital characteristics were well calibrated at this site during the period of analysis. Proprietary adjustments to the standard APACHE III model included the additional variable of pre-ICU treatment period and information about the institution size, teaching status, and region. In the case of the Princess Alexandra Hospital, a similar hospital model references the predictions to teaching hospitals of similar size in the Midwest region of the United States (C Alzola, personal communication, 1999).

Two control chart approaches to analysis of RA ICU and hospital mortality data are presented. Shewhart *p* charts (Figs. 1 and 2) were plotted by using semiannual observed mortality rates compared with APACHE III predicted mortality rates with control limits set at ±1.96 SD of the predicted mortality. Control limits are dependent on admission numbers and case-mix and are calculated from the APACHE III predicted mortality and an estimate of the SD of the predicted mortality rate. Appendix 1 provides references and the formulas for these estimates. In the design of the chart, limits were chosen for which, in a single observation period, the power to detect a doubling or halving of the odds of death would be estimated as ≥0.8.

The second chart analysis is a likelihood-based CUSUM chart to follow RA mortality. In Appendix 2, the CUSUM approach for continuous monitoring is described in detail. Appendix 3 presents an analysis of the performance and a description of the surveillance method in terms of run length characteristics.

The performance characteristics of the RA *p* chart and the RA CUSUM can be analyzed either in terms of the probability of signal during a period of analysis or the distributional features of the number of observations before a signal occurs, under both unchanged and changed odds ratios (OR). The first approach, the probability of false alarm and power of the technique, is familiar. Appendix 1 presents the formulas for this analysis of the RA *p* chart. Appendix 3 describes the analysis of the run length characteristics of the RA CUSUM. When the RA performance has not changed, the false-alarm run length is somewhat analogous to a measure of specificity. Where the RA performance has changed, run length to signal can be thought of as similar to sensitivity.

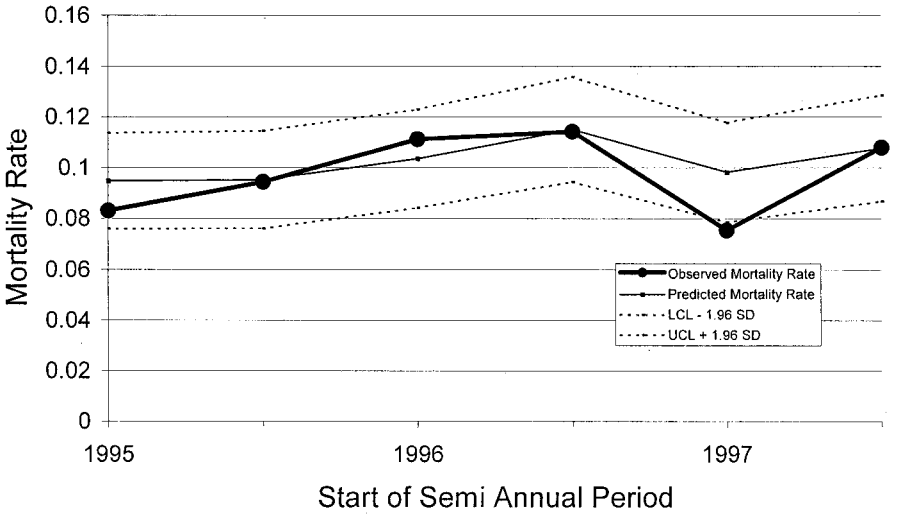

Figure 1. Risk-adjusted Shewhart *p*-chart of intensive care unit mortality for Princess Alexandra Hospital, 1995–1997. Modified *p* chart of semi-annual observed and predicted in-ICU mortality with ±1.96 SD control limits. Expected in-hospital mortality is estimated by the Acute Physiology and Chronic Health Evaluation III model, adjusted for hospital characteristics.

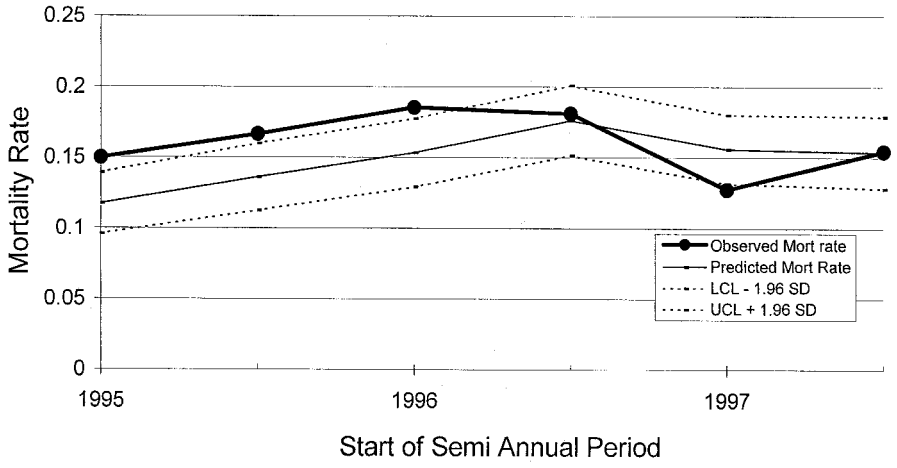

Figure 2. Risk-adjusted Shewhart *p*-chart of intensive care unit mortality for Princess Alexandra Hospital, 1995–1997. Modified *p* chart of semi-annual observed and predicted in-hospital mortality with ±1.96 SD control limits. Expected in-hospital mortality is estimated by Acute Physiology and Chronic Health Evaluation III model, adjusted for hospital characteristics.

RESULTS

The RA Shewhart *p* charts plotting semiannual mortality rates for in-ICU mortality (Fig. 1) and in-hospital mortality (Fig. 2) are plotted for the 3-yr period 1995–1997.

In Figure 1, the observed in-ICU mortality was close to predicted, except for

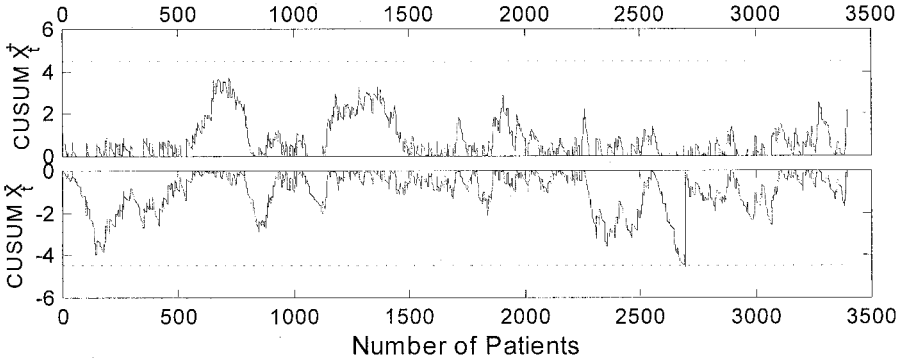

Figure 3. Risk-adjusted cumulative sum (*CUSUM*) of intensive care unit mortality for Princess Alexandra Hospital, 1995–1997. Expected in-hospital mortality is estimated by the Acute Physiology and Chronic Health Evaluation III model, adjusted for hospital characteristics. The control limit h is set at ±4.5.

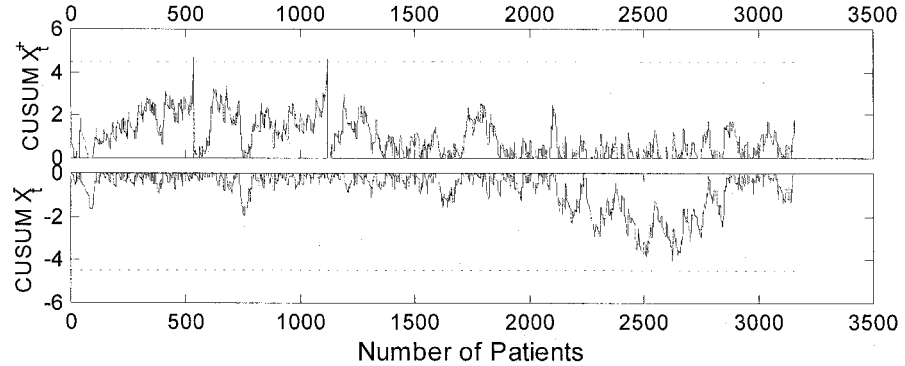

Figure 4. Risk-adjusted cumulative sum (*CUSUM*) of intensive care unit mortality for Princess Alexandra Hospital, 1995–1997. Expected in-hospital mortality is estimated by the Acute Physiology and Chronic Health Evaluation III model, adjusted for hospital characteristics. The control limit h is set at ± 4.5

the first half of 1997, during which the observed mortality rate decreased below the lower control limit of the predicted mortality. In Figure 2, the observed in-hospital mortality was higher than expected, with each of the first three review period rates falling beyond the upper 1.96 SD control limit.

The data are presented by using two-sided RA CUSUM charts for the ICU (Fig. 3) and the hospital outcomes (Fig. 4). In each, the upper CUSUM was designed to detect a doubling of the odds of death ($OR_A = 2$), and the lower chart was designed to detect a halving of the odds of death ($OR_A = 0.5$). When the CUSUM crosses the control limit (h^+ or h^-), enough evidence of a change in the ICU mortality rate has accumulated, and the CUSUM "signals." At this point, the CUSUM is reset to zero, and monitoring is resumed.

Figure 3 shows the variation around the baseline, with no accumulation of evidence that the odds of an ICU death are either substantially higher or lower than expected with the APACHE III ICU mortality model, in the first 2,600 patients. However, by admission number 2695, the CUSUM decreased below h^-. Either this signal was a random statistical event or the odds of in-ICU death were below predicted. The CUSUM was reset to zero, and there were no further signals.

Figure 4 presents a different pattern. The hospital RA CUSUM signaled an increased mortality by patient 533, when the upper control limit h^+ was exceeded. The upper RA CUSUM was reset to zero, and the upper control limit was exceeded again at patient 1119. It is reasonable to conclude that this represents increased odds of RA hospital mortality during this period. However, in a prospective application, a signal would prompt action, and subsequent signals would be interpreted in the context of any actions taken.

DISCUSSION AND CONCLUSIONS

RA control charting allows continuous monitoring of ICU outcomes while accounting for dynamic case-mix. Two different approaches are compared: the RA *p* chart is a summary charting technique, and the RA CUSUM is a sequential charting technique. This case series presents charts with signals of increased RA mortality exceeding APACHE III predictions and a signal representing random variation, or mortality decreasing below predicted.

The RA *p* chart analyses compare an observed mortality rate with the predicted mortality rate. In this example, by using ±1.96 SD control limits and semiannual reviews, the design of the RA *p* chart corresponds for a single observation period to the detection of a doubling of the odds of death (OR = 2) with a power of 0.98–0.99 for in-ICU mortality and 0.93–0.98 for in-hospital mortality. For a smaller increase in odds of death, say, OR = 1.5, the corresponding power estimates are in the ranges of 0.65–0.76 for in-ICU mortality and 0.45–0.59 for in-hospital mortality. The RA *p* chart will demonstrate large differences in RA performance but may be insensitive to small differences in RA mortality.

However, with six observation periods, the probability of one or more false alarms, where the probability of death is accurately predicted by the RA tool and where control limits are set at ±1.96 SD, is $1 - 0.95^6 = 0.26$. A false alarm would occur, on average, approximately every 10 yrs. It is very likely that the three signals of increased in-hospital mortality during the initial three semiannual reviews represent true increased odds of death relative to the APACHE III predictions. The signals of mortality below predicted in both the in-ICU and in-hospital mortality charts may be either chance events or real differences.

The RA *p* chart can be designed prospectively to meet meaningful performance specifications. For example, to detect a doubling of odds of in-ICU death with $\alpha = 0.05$ and $1 - \beta = 0.9$, given the case-mix of the sample, approximately 300 cases would be required (~14 wks).

With the design of the RA CUSUM, the average run length to signal, where there are no changes in the odds of in-ICU death, is 7150 admissions; that of in-hospital death is 5400 admissions. This is, on average, a false signal of increased RA mortality every 6.3 and 5.1 yrs, respectively. We conclude, again, that the signals of observed in-hospital mortality exceeding predicted are probably true-positive signals. The significance of the

signals of mortality less than predicted is unclear.

The choice of a specific control limit is a management decision, and chart design balances the risks of false signals and of not detecting a change in RA performance. Figure 5 in Appendix 3 shows the relationship between average run length and the changed OR in the context of in-ICU mortality and the specific control limit, $h = 4.5$. For example, the average run length to signal of the RA CUSUM when the OR is 2 is 125 cases (~5 wks).

The RA CUSUM may detect real changes sooner than the RA *p* chart, simply because the *p* chart cannot signal before the end of each observation period. The RA *p* chart requires arbitrary, defined periods of analysis and so may miss or delay a signal if changes occur halfway through a monitoring period. A characteristic of monitoring outcomes for quality management, compared with controlled interventional studies, is that the timing of changes may not be controlled or predicted in advance, or multiple, staged changes may occur. In contrast, the RA CUSUM can signal irrespective of the timing of the change in odds of death relative to the monitoring period if the increased OR is sustained. After signaling, the chart will resume accumulating evidence.

To interpret the significance of signals on RA control charts, we must consider the possibilities of random variation, systematic problems with model fit, and clinically relevant changes in the quality of care. Qualitative plots of observed minus expected outcomes do not account for the effects of random variation. Such plots have been used for cardiac surgical monitoring (12–14) but do not specify how much random variation in the plot is expected. Sherlaw-Johnson et al. (15) provide nested prediction intervals that have been applied to tracking cumulative mortality from myocardial infarction (16). However, the accumulating probability of false alarm with repeated analysis makes continuous monitoring of outcomes difficult to interpret.

In the analysis presented, either random variation or a nonsustained reduction in the odds of death would account for a lower than predicted mortality in the later part of the in-ICU and in-hospital series. However, the longer period of an apparent higher than predicted hospital mortality, with repeat signals, is less likely to be due to chance.

These techniques are as much a form of continuous assessment of RA tool calibration as of the clinical process of care. Where a change is signaled, either the model fit or the clinical milieu may have changed. It is mathematically not possible to separate the two possibilities, and, clearly, the model fit is dependent, among other things, on the clinical process. Changes in data collection (17, 18), patient case-mix (19), admission lead time (20), discharge practices (21), observed mortality rate (22), rule interpretations, and transcription ambiguities (23) all can potentially interfere with ICU mortality prediction systems, including the APACHE III system. It is essential that the models for RA be validated on site and that model or data failure always be considered as a possible cause for a signal. This application uses a validated generalization of an existing model (11), but the "black box" nature of the APACHE III system makes it difficult to analyze the role of the RA model when a signal is detected. Recalibration of an existing ICU outcome model (21, 24, 25) or application of locally developed models would be equally appropriate and more transparent.

In this analysis, the data collection and application of the RA tool were not altered, and a change in survival is possible. An improvement in senior staffing was a discrete milestone that occurred around admission number 1100. An ongoing multidisciplinary, evidence-based review of all aspects of patient care followed and continues, similar to guidelines recently proposed independently elsewhere (26). No further signals of increased RA mortality were seen in this analysis.

Considering these caveats, RA outcome monitoring is proposed as a method of quality management. General ICUs seldom have the caseload and never have the resources to offer controlled prospective evaluation of incremental change. Cardiac surgery is a common group of procedures, yet a power analysis of 43 Veterans Affairs hospitals in the United States (1987–1992) (27) found that only one institution had enough cases to potentially detect a doubling of surgical mortality in a 6-mo period ($\alpha = 0.05$; $1 - \beta = 0.8$). A typical, general ICU has a myriad of diagnoses. APACHE III has 95 disease groups and more than 250 diagnoses. Well-designed controlled studies can demonstrate the effectiveness of a clinical intervention under test conditions, but application to local clinical practice conditions may be a concern (6). Collecting similar patients leads to delays in data accumulation, with therapeutic creep and continual change rendering historical findings obsolete. Excess mortality during case collection may be an unacceptable price to pay for specificity. Timely recognition may support further improvements or prevent unnecessary increased mortality.

Continuous monitoring of RA outcomes allows a broad evaluation of a clinical milieu. These are not tools to assess a change that affects only a small number

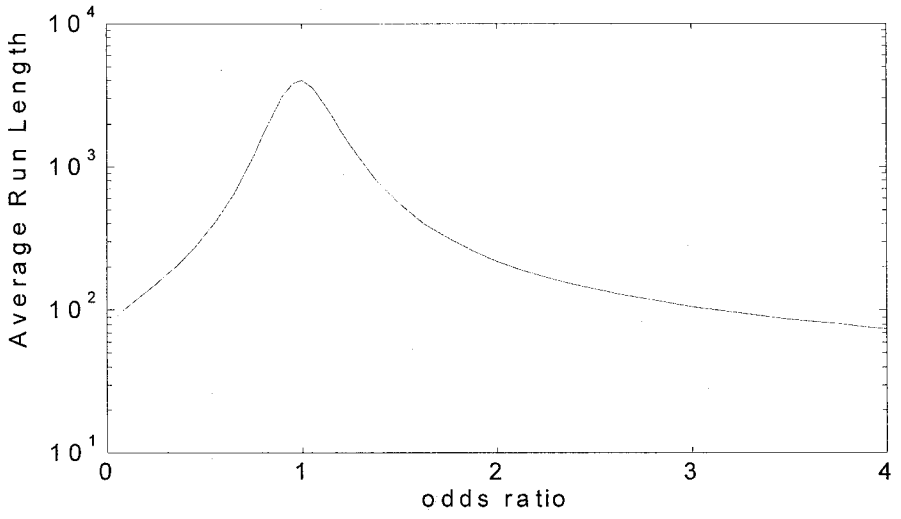

Figure 5. Average run length of cumulative sum for changes in the odds ratio of risk-adjusted mortality. The distribution of average run length is based on the case mix of the Princess Alexandra Hospital intensive care unit and the choice of *h* at ±4.5. An odds ratio of 1 gives an average run length of 7400, whereas a doubling or halving of the odds ratio should signal by 125 patients on average.

This article presents the first continuous monitoring technique of outcomes in an intensive care unit with risk-adjusted control charting to track local risk-adjusted mortality performance.

of patients; they are more suited to monitoring changes that affect all patients or the entire clinical process. RA monitoring is not a substitute for rigorous controlled evaluation, but, then, controlled studies are not necessarily practical for the evaluation of a culture of innovation, evolution, and incremental change.

This article presents the first continuous monitoring technique of outcomes in ICU with RA control charting to track local RA mortality performance. Medical practice is an evolutionary process, which, we assume, moves toward improved patient outcomes. The methods are proposed so that we may contemporaneously learn the most from our patient data, rather than be judged by others in retrospect.

REFERENCES

1. Benneyan JC: Use and interpretation of statistical quality control charts. *Int J Qual Health Care* 1998; 10:69–73
2. Parry BR, Williams SM: Competency and the colonoscopist: A learning curve. *Aust NZ J Surg* 1991; 61:419–422
3. Williams SM, Parry BR, Schlup MMT: Quality control: An application of the CUSUM. *BMJ* 1992; 304:1359–1361
4. Schlup MTM, Williams SM, Barbezat GO: ERCP: A review of technical competency and workload in a small unit. *Gastrointest Endosc* 1997; 46:48–52
5. de Leval MR, Francois K, Bull C, et al: Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations. *J Thorac Cardiovasc Surg* 1994; 107:914–924
6. Levett J, Carey R: Measuring for improvement: Toyota to thoracic surgery. *Ann Thorac Surg* 1999; 68:353–358
7. Kennett R, Zacks S: Chapter 11: Advanced methods of statistical process control. *In:* Modern Industrial Statistics: Design and Control of Quality and Reliability. Belmont, CA, Duxbury Press, 1998, pp 360–407
8. Steiner S, Cook R, Farewell V, et al: Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; 1:441–452
9. Steiner S, Cook R, Farewell V: Risk adjusted monitoring of surgical outcomes. *Med Decis Making* 2001; 21:163–169
10. Knaus WA, Wagner DP, Draper EA, et al: The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest* 1991; 100: 1619–1636
11. Cook D: Performance of APACHE III models in an Australian ICU. *Chest* 2000; 118: 1732–1738
12. Lovegrove J, Valencia O, Treasure T, et al: Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997; 350:1128–1130
13. Poloniecki J, Valencia O, Littlejohns P: Cumulative risk adjusted mortality chart for detecting changes in death rate: Observational study of heart surgery. *BMJ* 1998; 316: 1697–1700
14. Poloniecki J. Letter. *BMJ* 1998; 317:1453
15. Sherlaw-Johnson C, Lovegrove J, Treasure T, et al: Likely variations in perioperative mortality associated with cardiac surgery: When does high mortality reflect bad practice? *Heart* 2000; 84:79–82
16. 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. *BMJ* 2001; 323:324–327
17. Bosman RJ, Oudemans van Straaten HM, Zandstra DF: The use of intensive care information systems alters outcome prediction. *Intensive Care Med* 1998; 24:953–958
18. Chen D, Martin CM, Morrisin TL, et al: Interobserver variability in data collection of the APACHE II score in teaching and community hospitals. *Crit Care Med* 1999; 27: 1999–2004
19. Murphy-Filkins R, Teres D, Lemeshow S, et al: Effect of changing patient mix on the performance of an intensive care unit severity-of-illness model: How to distinguish a general from a specialty intensive care unit. *Crit Care Med* 1996; 24:1968–1973
20. Dragsted L, Jorgensen J, Jensen N, et al: Interhospital comparisons of patient outcome from intensive care: Importance of leadtime bias. *Crit Care Med* 1989; 17:418–422
21. Sirio CA, Shepardson LB, Rotondi AJ, et al: Community-wide assessment of intensive care outcomes using a physiologically based prognostic measure. *Chest* 1999; 115:793–801
22. Zhu HP, Lemeshow S, Hosmer DW, et al: Factors affecting the performance of the models in the Mortality Probability Model II system and strategies of customization: A simulation study. *Crit Care Med* 1996; 24:57–63
23. Fery-Lemonnier E, Landais P, Loirat P, et al: Evaluation of severity scoring systems in ICUs—Translation, conversion and definition ambiguities as a source of interobserver variability in APACHE II, SAPS and OSF. *Intensive Care Med* 1995; 21:356–360
24. Le Gall J, Lemeshow S, Leleu G, et al: Customised probability models for early severe sepsis in adult intensive care. *JAMA* 1995; 273:644–650
25. Apolone G, Bertolini G, D'Amico R, et al: The performance of SAPS II in a cohort of patients admitted to Italian ICUs: Results from GiViTi. *Intensive Care Med* 1996; 33: 1368–1378
26. Randolf A, Pronovost P: Reorganizing the delivery of intensive care could improve efficiency and save lives. *J Eval Clin Pract* 2002; 8:1–8
27. Marshall G, Shroyer LW, Grover FL, et al: Time series monitors of outcome: A new dimension for measuring quality of care. *Med Care* 1998; 36:348–356
28. Montgomery D. Introduction to Statistical Quality Control. Third Edition. New York, Wiley, 1996
29. Kennett R, Zacks S: Modern Industrial Statistics: Design and Control of Quality and Reliability. Belmont, CA, Duxbury Press, 1998
30. Flora J: A method for comparing survival of burn patients to a standard survival curve. *J Trauma* 1978; 18:701–705
31. Steiner SH, Cook RJ, Farewell VT: Monitoring paired binary surgical outcomes using cumulative sum charts. *Stat Med* 1999; 18: 69–86
32. Knaus WA, Draper EA, Wagner DP, et al: APACHE II: A severity of disease classification system. *Crit Care Med* 1985; 13: 818–829
33. Le Gall J, Lemeshow S, Saulnier F: A new simplified acute physiology score (SAPS II) based on a European/North American multicentre study. *JAMA* 1993; 270:2957–2963
34. Lemeshow S, Teres D, Klar J, et al: Mortality probability models (MPM II) based on an international cohort of intensive care unit patients. *JAMA* 1993; 270:2478–2486
35. Lemeshow S, Klar J, Teres D, et al: Mortality probability models for patients in the ICU for 48 or 72 hours: A prospective multicenter study. *Crit Care Med* 1994; 22:1351–1358

APPENDIX 1

Control Limits and Power Analysis of Shewhart p *Chart.* The modified *p* chart plots observed mortality rates, predicted mortality rates, and estimated control limits. In situations in which the cases have a differing expectation of mortality, the usual formulas for calculation of control limits (28, 29) are modified.

Control limits are dependent on admission numbers and case mix and are calculated from the Acute Physiology and Chronic Health Evaluation (APACHE) III mortality predictions. An estimate is

made of the SD of the predicted mortality rate (15, 30). This leads to

$$CL = \frac{1}{n}\left(\sum_{i=1}^{n} p_i \pm K \sqrt{\sum_{i=1}^{n} p_i(1-p_i)} \right) \qquad [1]$$

where: *CL* are the control limits, n is the number of cases in each block of admissions, indexed by i, p_i is the predicted risk of death, K is the number of standard deviations chosen for control limits, and ± 1.96 SD is equivalent to an alpha error of .05 for a single observation.

More accurate characterization of the distribution of the observed mortality for the Princess Alexandra Hospital data set, using simulations or a Markov, iterative approach demonstrates that this estimate of the control limits is good to ± 2 SD, but it becomes less reliable for decision thresholds beyond. In monitoring health outcomes, particularly mortality, control limits beyond of ± 2 SD may demand an inappropriately high specificity before prompting review and action. Also, given the performance of the RA models, accuracy beyond 2 SD may be unrealistic.

The power of the analysis to detect changes in the RA mortality in any single observation period can be estimated given case mix, block grouping size, and an alternative hypothesis to define risk of death. This calculation may be useful in the design of the chart.

$$Power = \Phi\left\{ \frac{-K\sqrt{\sum_{i=1}^{n} p_i(1-p_i)} + \sum_{i=1}^{n}(p_i - Q_i)}{\sqrt{\sum_{i=1}^{n} Q_i(1-Q_i)}} \right\} + \left[1 - \Phi\left\{ \frac{K\sqrt{\sum_{i=1}^{n} p_i(1-p_i)} + \sum_{i=1}^{n}(p_i - Q_i)}{\sqrt{\sum_{i=1}^{n} Q_i(1-Q_i)}} \right\} \right] \qquad [2]$$

Where $\Phi(x)$ is the cumulative normal distribution for the value x and Q_i is the alternative risk of death. In this series, it is based on an altered odds ratio (OR_A). If

$$OR_A = \frac{Q_i/1 - Q_i}{p_i/1 - p_i} \qquad [3]$$

then

$$Q_i = \frac{OR_A \cdot p_i}{1 - p_i + OR_A \cdot p_i} \qquad [4]$$

APPENDIX 2

Risk-adjusted Cumulative Sum Procedure. The development of the risk-adjusted cumulative sum (RA CUSUM) approach has been described in detail elsewhere (8, 9, 31). A statistic, X_t, accumulates the score for all patients from the start of monitoring until the point of observation. Mathematically, an RA CUSUM chart is created by plotting X_t vs. patient number t, where

$$X_t = \max(0, X_{t-1} + w_t) \qquad [5]$$

and

$$t = 1,2,3\ldots$$
$$X_0 = 0\,,$$

The patient's score (w_t) depends on three factors: the patient's estimated risk of death, p_t, the patients outcome y_t, where $y_t = 0$ for a survivor and $y_t = 1$ for a death, and OR_A, defining the alternative level of performance to be detected. The risk, p_t, is provided by an RA model. The APACHE III model adjusted for hospital characteristics is used. Any existing model such as APACHE II (32), SAPS II (33), or the MPM (34, 35) series could be used (recalibrated if the fit were inadequate), or a site specific logistic regression model could be fitted to sample data and be validated. For "near real time" monitoring of RA mortality, an intensive care unit (ICU) model with a 30-day survival end point would, ideally, permit analysis to be only 30 days in arrears.

Because each patient's mortality risk may vary, OR_A is specified as an OR. In this example, we design a chart to detect a doubling in the odds of failure, so we set $OR_A = 2$. An increase in the OR to 2 is equivalent to an increase in the overall ICU mortality rate from the current 9.9% (16% in hospital) to 18% (27.6% in hospital).

The score, (w_t) for each patient is derived using a log likelihood ratio. It is given by the logarithm of the ratio of the probability of the outcome observed, under the alternative hypothesis of interest, which is defined through OR_A and p_t, to the probability under the currently estimated risk, defined as p_t if the outcome of interest is death and $(1 - p_t)$ if the outcome is survival.

$$w_t = \begin{cases} log\left[\dfrac{OR_A}{(1 - p_t + OR_A p_t)} \right] & \text{if } y_t = 1 \text{ (i.e., patient } t \text{ dies)} \\ \log\left[\dfrac{1}{1 - p_t + OR_A p_t} \right] & \text{if } y_t = 0 \text{ (i.e., patient } t \text{ survives)} \end{cases} \qquad [6]$$

The CUSUM can be regarded as a formal sequential procedure for testing the null hypothesis: H_0: OR = 1, vs. the alternative hypothesis, H_A: OR = OR_A.

When designing a chart to detect increases in the ICU mortality rate (i.e., ORA > 1), the w_t associated with mortality are positive, whereas successes receive a negative score. For example, if $OR_A = 2$, then a patient who has a predicted ICU mortality risk of 1% would contribute a score equal to 0.68 if her or she dies and -0.01 if he or she lives. A patient who has a predicted mortality risk of 30% would contribute a score equal to 0.43 if he or she dies and -0.26 if he or she lives. The "penalty" for a death of a low-risk patient is more severe than for a death of a higher-risk patient.

To complete the CUSUM design, we must choose a value for the control limit (h). For each of the CUSUMs, the choice of h involves an inherent tradeoff based on the in-control and out-of-control average run length of the proposed procedure (Appendix 3). For the CUSUM designed to detect increases (decreases) in the ICU mortality rate, setting the control limit h at 4.5 (-4.5) gives an average run length of around 7,150 patients when the OR is unchanged.

For presentation on the same diagram of two CUSUM procedures on the same data series, we have used the convention that if $OR_A < 1$,

$$X_t = \min(0, X_{t-1} - w_t) \qquad [7]$$

and the CUSUM accumulates a negative value. The procedure signals when X_t falls below $h^- = -4.5$

APPENDIX 3

Analysis of Performance of the RA CUSUM Monitoring Scheme. A CUSUM is designed to continually monitor the ICU performance until a signal occurs. The

procedure will, theoretically, always eventually signal because the testing (monitoring) is continued indefinitely. The number of patients until the CUSUM first exceeds the control limit is called the run length of a CUSUM. If the ICU mortality rate has not changed, a signal will represent a false alarm, and long run lengths are desirable. If the mortality has increased substantially, short run lengths are desirable.

The run length is a random variable whose distribution represents all the possible runs that may arise given a particular overall mortality rate, patient mix, and the effects of chance. Thus, the average run length of the CUSUM when the mortality rate has not changed is somewhat similar to the type I error rate of a traditional statistical test. Likewise, the average run length of the CUSUM when the mortality rate has increased substantially is somewhat analogous to the power of a traditional statistical test. Determining the average run length of a CUSUM at the design stage is computational intensive because it is based on all possible outcomes for a long series of patients, but it may be closely approximated (31). Selection of h represents a trade-off between false alarms and delay in detection of a true change in the odds of death.

To quantify the ability of the CUSUM procedure to quickly detect increases in the odds of death, Figure 5 shows a plot of the average run length against a measure of the in-ICU performance given in terms of the RA odds of mortality. We use the patient mix, in terms of APACHE III risk of death estimates in the current data set, to estimate these distributions. We see in Figure 5 that substantial changes in the OR result in a rapid decrease in the expected run length of the CUSUM.