Review of the application of risk-adjusted charts to analyse mortality outcomes in critical care

David A Cook, Graeme Duke, Graeme K Hart, David Pilcher and Daniel Mullany

ABSTRACT

This review describes the methods for displaying risk-adjusted mortality data for critical care units. Two applications are considered. The comparison within a cohort of risk-adjusted mortality performance uses standardised mortality ratios (SMRs), league tables, caterpillar plots and funnel plots. Monitoring of risk-adjusted performance over time is considered using SMRs, risk-adjusted *p* (RAP), observed minus expected outcome (VLAD), risk-adjusted cumulative sum (RACUSUM), risk-adjusted sequential probability ratio test (RASPRT), and risk-adjusted exponentially weighted moving average (RAEWMA) charts. Examples of the charts are provided, and calculation of the statistics and design of the charts are described in the Appendix.

This overview is an introduction to the use of risk-adjustment methods to track mortality rates. The importance of model performance and relevance of the risk-adjustment models is emphasised. The relative merits of different methods are discussed. Risk-adjusted monitoring plays a role in the context of a holistic quality development strategy. The importance of a planned approach to response and intervention is stressed.

Crit Care Resusc 2008; 10: 239–251

Counting survivors and deaths provides one measure of the effectiveness of health care interventions and complements other assessments of unit-based performance, including process-of-care key performance indicators and regular audit of process compliance and index events. However, several factors other than variations in the quality of care influence the observed number of deaths or mortality rate. Major contributors include casemix, severity of illness and physiological reserve of patients, lead time to admission,[1] and discharge practices.[2] Many other unmeasured factors can potentially confound comparisons between the raw mortality rates. Understanding these issues is an active area of research.

One response is to monitor patient outcomes by measuring the mortality rate in a defined group of similar or matched patients, under strictly defined conditions. This is the approach taken in empirical research in controlled trials. It constrains the natural variation in otherwise unselected patient populations, but limits the proportion of critical care patients who contribute to an analysis. An alternative approach is to statistically control for measurable variations in patient casemix, severity of illness and risk of death. Where these techniques are incorporated into analysis of death rate, the analysis is considered to be *risk-adjusted.*

Risk-adjustment models relevant to the critical care population (eg, APACHE II,[3] APACHE III,[4] and UK APACHE models[5]) can be used to draw limited inferences and comparisons over time and within populations.[6-8] The risk-adjustment model, and in effect the dataset on which the model was developed and calibrated, provide a target or benchmark by which performance is measured.

It is not possible to statistically control for all the factors that could contribute to a patient's risk of death. No model is perfect, and some of the variation in observed mortality rates will be due to unmeasured factors. These include variations in data collection methods and data quality, variations due to imperfect model predictions, and unit- or hospital-specific variations. Some are completely random influences. Each of these potential sources of error can be modelled, either as a random variation (acting in either direction), or a bias (tending to act in one direction). While these cannot be eliminated, steps can be made toward quantifying the possible random variations and known model biases during the risk-adjustment process. Estimating the magnitude of these errors permits analysis of the amount that observed differences in performance could be due to factors related to the individual unit or hospital quality of care.

The accuracy and performance of a model in predicting a patient's risk of death can be quantified by statistical means. As a generality, the more accurately and reproducibly a model estimates the risks of death (or any other outcome of interest), the more its potential usefulness as a risk-adjustment tool. There are two important aspects of model performance:[9] calibration and discrimination.

Calibration is the accuracy of a model in estimating a patient's real risk of death. Inaccurate calibration introduces a bias in the estimates. For example, a model that overestimates patients' risk of death will be biased, and under these

circumstances most units will observe outcomes that are better than predicted. The extent of bias depends on the calibration of the model across the range of diagnostic groups, severity of illness spectrum and casemix in the sample under analysis.

Discrimination is the ability of a model to allocate a higher probability of death to a patient who ultimately dies than to a patient who survives. If this cannot be done reliably, then the model is little better than chance, irrespective of its calibration.

Intensive care models perform well, but this prediction accuracy tends to decline with time. For example, the APACHE III model in Australia has performed well in the past,[10,11] with little bias and high discrimination. However, the APACHE III-J developed in the early 1990s is increasingly tending toward overestimating the risk of death in Australian intensive care units (unpublished data from ANZICS CORE database), while still maintaining high discrimination. This phenomenon could be due to improvement in patient survival from critical illness over time, and has been referred to as "model fade" by the APACHE development team.[12]

An approach to this problem of model fade is to routinely check the model fit, and to re-calibrate the model as required to reflect current practices and outcomes. The Cerner Corporation (supplier of the APACHE system) refits the APACHE model from time to time and recently released the APACHE IV update as part of its proprietary database solution.[13]

Although data quality and robust risk adjustment are the most important components of any monitoring process, the methods by which data are analysed and presented contribute to interpretation of the analysis. This review summarises several available methods, and focuses primarily on the role of process control charts. These methods for intensive care data analysis are under consideration by data monitoring groups, such as ANZICS CORE and the Victorian Data Intensive Care Data Review Committee (VICDRC) of the Department of Human Services, Victoria. The application of control charts in health care monitoring[14] and risk-adjusted methods[15,16] are reviewed in more detail elsewhere.

Types of risk-adjusted analysis

Three applications of continuous risk-adjusted monitoring and analysis are described.

Cross-sectional risk-adjusted charting

Typically, this approach is used to analyse historical data and provide comparisons between logical cohorts of individuals, clinical units or institutions, using aggregated or grouped samples of data.

Examples of these analytical techniques include:

- the standardised mortality ratio (SMR);
- the funnel plot; and
- league tables.

The implicit comparison is between the relative performance among the cohort of participating units, and against an external target or benchmark performance derived from the risk-adjustment model predictions.

Longitudinal or sequential risk-adjusted charting

This approach displays sequential episodes of patient care over time to detect trends or patterns in risk-adjusted outcomes. It concentrates on detecting runs of improved or deteriorating mortality performance relative to the external standard based on the risk-adjustment tool.

Examples include:

- risk-adjusted *p*-charts (RAP charts);[8]
- risk-adjusted exponentially weighted moving average (RAEWMA) charts;[17] and
- various forms of cumulative sum (CUSUM) charts, such as:
  - observed minus expected values charts (variable life-adjusted display [VLAD][18,19] or cumulative risk-adjusted mortality [CRAM][20] charts);
  - risk-adjusted CUSUM (RACUSUM);[21,22] and
  - risk-adjusted sequential probability ratio test (RASPRT)[23,24] charts.

Analysis to improve risk-adjusted estimates of probability of death

This analysis emphasises accurate estimation of the probability of patient death. The model estimates are modified by a sequential analysis of the model predictions and observed outcomes. The purpose is to provide accurate estimates of probability of death, rather than comparisons of performance between institutions or over time. Charts of the difference between the observed and expected outcomes[20] and Bayesian methods[25] have been proposed.

In this review, we focus on the first two applications of risk-adjusted analyses.

Caveats, cautions and dangers

Cross-sectional and longitudinal displays of risk-adjusted patient outcomes have two features in common. They provide a simplified visual display of complex datasets, and enable comparison with a benchmark, usually provided by the risk-adjustment tool estimates. Another feature common to most is the addition of control limits. An observation of mortality rate situated within control limits is described as an "inlier", where the difference between the observation and the prediction is within the

realm of a reasonable chance occurrence. In contrast, an observation outside these control limits is termed an "outlier", as the observation is likely not due to a chance occurrence. The statistical "outlier" may or may not be a clinical "outlier" in terms of clinical practice. The quality of the data and the robustness of the risk-adjustment model are crucial to any analysis or display method. Missing or inaccurate data or poor performing models increase the risk of misleading results and can make the analysis uninterpretable.

From the preceding introductory discussion, several cautions and caveats must be made regarding risk-adjusted monitoring. These apply to all the analytical methods. Risk-adjusted analysis is ultimately an assessment of model fit on a sample of data. Conclusions from the risk-adjusted comparisons and analysis flag only whether the model fits the sample data. Enquiry into the causes of this should follow. One of all the possible causes is a variation in the quality of patient care. In our experience, deficiencies in quality of care and examples of excellent care can be, and are, identified by risk-adjusted analyses. However, many (or perhaps most) statistical anomalies are due to fluctuations in data quality and completeness, systematic errors in coding, or changes in discharge and referral practice.

Reports[24] of poor performing providers are usually retrospective analyses. Some of this is publication bias toward interesting examples. More importantly, our anecdotal experience has been that potentially real examples of diminished performance that have been detected cannot be published, due to confidentiality, data ownership and the great sensitivity of the material.

If the data are of poor quality or unreliable, meaningful analysis may not be possible. More importantly, poor data quality can lead to spurious and unwarranted conclusions about the clinical outcomes. Incomplete data, particularly where there is a bias toward collecting the deaths or survivors, will confound any analysis, irrespective of the method.

Poor model fit, even with accurate and complete data, will make conclusions very difficult. Recalibration of the current APACHE models to Australian data would be desirable before they can be used as more than a screening tool for outliers. A risk-adjustment model that is a poor discriminator will cause more false alarms in both directions unless pooled samples (RAP chart or serial SMR) or RAEWMA charts are used.

All methods require an accumulation of evidence, and, in the presence of a low predicted mortality rate or a small case load with few patients, any statistical analysis will lack power to demonstrate anomalies in a timely manner. Under these conditions, the variability within the data and inherent inaccuracies in the model predictions can be greater than any plausible clinical variation in outcomes caused by quality of care. The confidence intervals and control limits can indicate precision and expected variations of the observations.

Finally, most applications of risk-adjusted analysis are done by centralised data collection and collation organisations for the purpose of performance monitoring. The outcomes that are monitored (survival to ultimate hospital discharge) and the timeliness of data submission and analysis impose delays of up to months until the analysis can be reviewed. This is not an ideal situation for incorporating risk-adjusted analysis into practice improvement. The geographical and temporal distance between the events under scrutiny diminishes the importance and relevance of the final analysis. The ideal vehicle for risk-adjusted analysis is a well fitting model, with good quality and reliable data collected, submitted and compiled contemporaneous to the clinical episode, short-term outcomes (such as 28-day survival) and a flexible data analysis tool.

Examples of risk-adjusted analysis and charts

The following examples are of charts and analysis that have been used in critical care mortality monitoring and analysis. The Appendix shows the formulas for calculating and charting of these analyses.

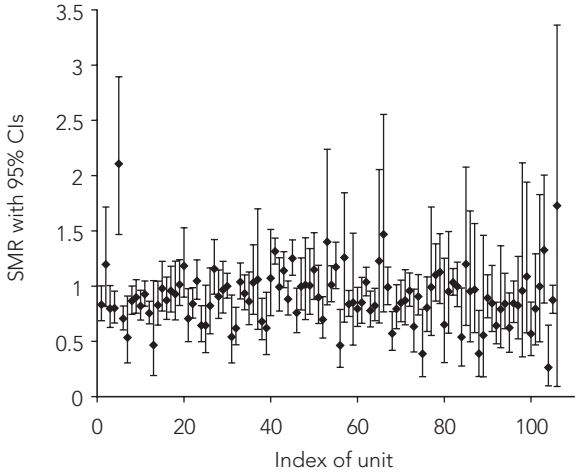
Charts for population or cross-sectional analysis

Standardised mortality ratio (SMR)

The SMR is a commonly used summary statistic and is presented in the familiar ANZICS CORE Adult Patient Database reports. It is the ratio of observed deaths to predicted deaths, or observed mortality rate to predicted mortality rate. A value above unity indicates more deaths than the reference model predicts, while a value less than unity represents fewer deaths than predicted. A confidence interval should always be included to demonstrate the precision of the SMR estimate. If the confidence intervals cross unity then there is inadequate statistical evidence of a difference from the benchmark.

Several ways to estimate the confidence intervals around the SMR estimate are described in the Appendix. The SMR provides a summary statistic for the sample that obscures possible differences in patient subgroups or seasonal variations. Sequential SMR values (eg, annually) provide some means to assess change over time. An example of an SMR analysis is shown in Figure 1.

Figure 1. Standardised mortality ratios (SMRs) for 106 Australian and New Zealand ICUs in 2004*



** Based on de-identified data submitted to the ANZICS database, with APACHE III-J risk adjustment.*

League tables

League tables provide a summary representation of risk-adjusted performance. One approach is to compare units, providers or institutions ranked according to the estimate of the SMR, giving a rank order of performance.

The limitation[23] is that league tables present an apparent precision of ranking that is misleading and can be easily misrepresented. This occurs because imprecision in the predicted risks of death and statistical uncertainty around the observed risk of death create quite wide confidence intervals around each SMR. The rank given to each provider then has a great deal of uncertainty. Unfortunately, as the rank is an ordinal number, it carries a pejorative appearance of precision.

A league table can be most fairly presented as a caterpillar plot. Figure 2 shows a caterpillar plot of league rankings of providers by SMR with 95% confidence intervals around the SMR estimate. The imprecision and possible variation in SMR estimates show clearly that the rank of any unit with respect to the others is uncertain. Ranks and even quartile positions are not reliable indications of relative performance unless these positions are consistently and reproducibly achieved.

Funnel plots

The funnel plot[26] is a graph of mortality rate or SMR observations from a number of institutions plotted against the number of eligible cases or sample size for each provider during the period of analysis. As the sample size increases, so does the precision of the observations. The funnel plot is so named because of the funnel shape of the confidence intervals. The confidence interval is around the predicted risk of death, and is based on the size of the population. For a constant mortality rate, wide confidence intervals occur with small patient numbers, and narrower confidence intervals with large patient numbers.

Typically, the data from the cohort of providers are collected over the same time period. The SMR is a convenient risk-adjusted statistic to use for each unit, and is compared with a target SMR of unity, with prediction intervals that encompass 95% and 99% of expected values providing the funnel appearance. The relationship between the plotted SMR for the provider and the confidence intervals, the relationship with sample size, and the relative position with

Figure 2. Caterpillar plot of standardised mortality ratio (SMR) for the 106 ICUs in Figure 1, ranked by SMR estimate

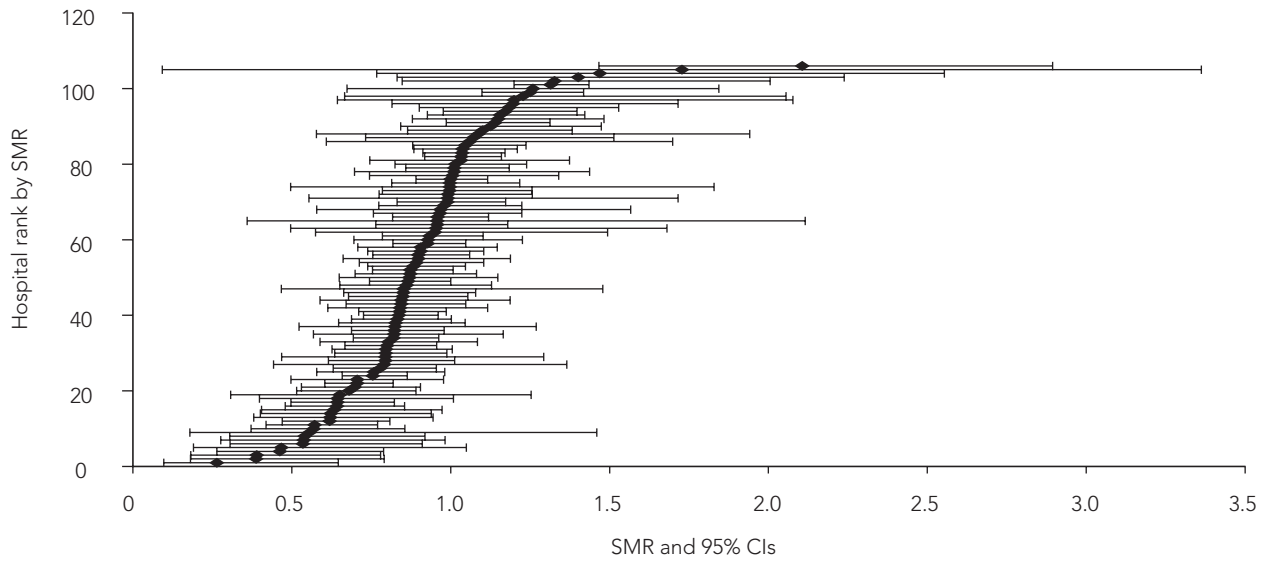
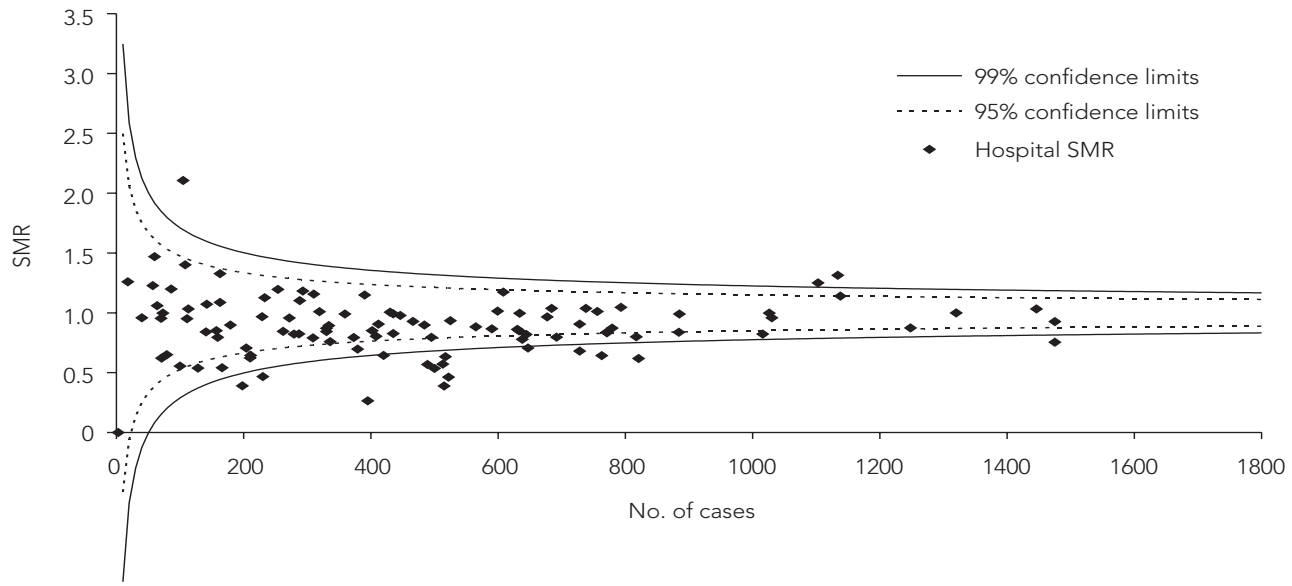


Figure 3. Funnel plot of standardised mortality ratio (SMR) for the 106 ICUs in Figure 1, by number of cases



respect to the other providers in the cohort are readily seen. Mortality rate after risk adjustment can be assessed as a potential high or low statistical anomaly (Figure 3).

The purpose of a funnel plot is to present mortality rate adjusted for risk of death, with a visual representation of the degree of uncertainty, related directly to sample size. To do this, control limits or the confidence intervals around the line SMR = 1 must be directly related to the sample size on the *y* axis. It is necessary to disregard casemix and mortality rate considerations for samples drawn from the population, and to assume a constant mortality rate across the samples.

The limitations of the funnel plot are similar to those of the SMR. This method underestimates the actual spread of SMRs that will be seen, and some adjustments for this have been recommended.[27] This may not be quite so important with mortality rates above 10% and those which are reasonably constant among providers, and when models with good discrimination are used.

Charts for longitudinal and sequential analysis

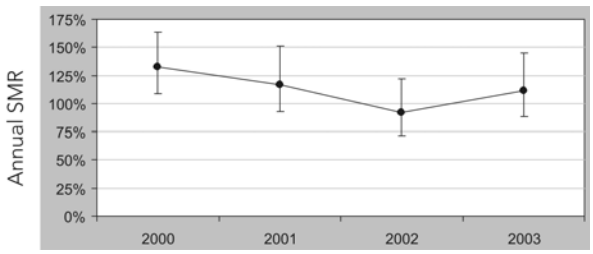
Sequential process control charts provide a longitudinal analysis of sequential data and are available in several different formats. Most are used for individual patient observations, but RAP charts and sequential SMRs are often used to group a consecutive series of patients. Grouped sequential observations can be used for any of the other charts described.

The strength of sequential control charts is their ability to identify (and quantify) emerging data trends quickly. They are useful as a screening tool and aid to real-time monitoring of clinical performance. Like the methods mentioned above, they can be used to identify improvements in care (better performer), not just deteriorations in care (poor performer).

All formats of sequential process control charts display the cumulative or progressive sum of the differences in the observed and expected outcomes. The abscissa is in sequential patient order of presentation, which can be mapped to a calendar date.

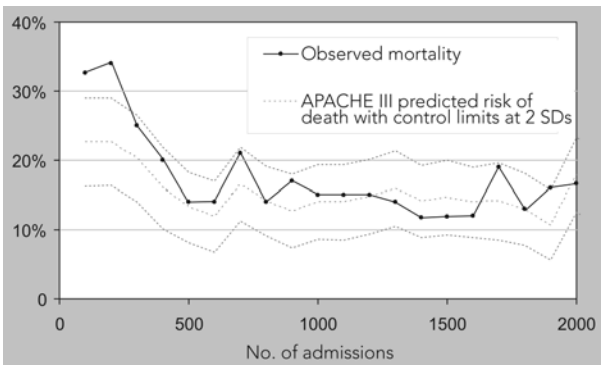
Design of the charts and interpretation of any signals or statistical anomalies requires an understanding of the distribution of run lengths expected under conditions where the risk of death is as predicted with the model, and under a range of clinically plausible altered risks of death. A run length is the number of sequential cases needed to accumulate enough evidence to elicit a signal, either correctly or falsely. The average run length (ARL) is used as a summary statistic of the expected behaviour of the monitoring. For example, the ARL of a RACUSUM chart monitoring a context where the odds of death are 1 (ie, the model predicts the outcomes accurately) will be a large number. This means that signals or alarms will seldom occur, and are due entirely to chance, so false alarms are unlikely. However, it is desirable that the ARL is a small number, where there is a clinically significant change in observed outcomes compared with expected outcomes. Thus alarms or signals should occur quickly to minimise the delay in detecting a real difference. Charts are designed to balance these two detection considerations between true signals and false alarms. Characterisation of run length distribution is complex, and often requires a simulation or Markov Chain Monte Carlo methods to estimate the distributions, and thus the ARLs, across a range of scenarios.

Figure 4. Annual APACHE III-standardised mortality ratio (SMR) for a hypothetical hospital (QPDH), 2000–2003



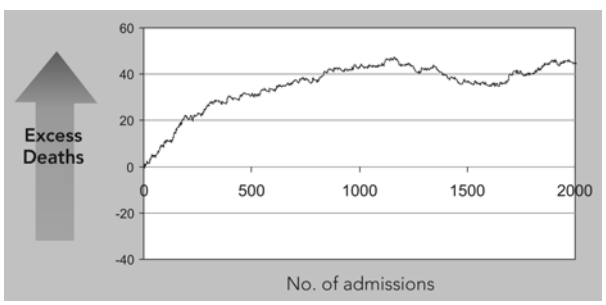
500 patients were admitted per year to Queen Priscilla's Desert Hospital (QPDH). The annual SMR and associated 95% CIs are shown. Where the confidence interval extends across 100%, mortality is not significantly different from that predicted by APACHE III.

Figure 5. Risk-adjusted *p* (RAP) chart for QPDH, 2000–2003



Data are presented in sequential groups of 100 patients. For each group, observed mortality is plotted with APACHE III predicted mortality and its associated control limits.

Figure 6. Observed minus expected values (VLAD) chart for QPDH, 2000–2003



The y axis represents additional deaths over the number predicted by APACHE III. Negative numbers (ie, below the x axis) represent more survivors than predicted by APACHE III. No control limits are shown, and therefore it is not clear whether the excess deaths are statistically significant.

The propensity for false alarms and the ability to detect true differences can also be approximated by considering Type I and Type II errors for the RAP charts and the RASPRT chart. Strictly, these are best assessed by ARL analysis.

Serial standardised mortality ratios

A familiar representation of risk-adjusted data is a display of the SMRs with confidence intervals for consecutive time periods. Figure 4 shows the SMRs from simulated data for the hypothetical Queen Priscilla's Desert Hospital (QPDH).

Risk-adjusted p (RAP) charts

The RAP chart plots the observed mortality rate for a group of patients compared with a predicted mortality rate for the sample, with control limits around the estimate calculated using the risks of mortality of patients in the sample. Risk-adjusted $\bar{x}$ charts[28] and *p* charts[29] have been proposed by Alemi and colleagues using the *t* distribution to calculate control limits. The ICU application has large sample sizes, so it is acceptable to adapt the RAP chart to monitoring of ICU risk-adjusted mortality, using a normal approximation to calculate the control limits. For this method, the population must be divided into discrete cohorts (eg, 100 consecutive patients), and the rate for each cohort is compared with the predicted rate. For administrative ease, the patient samples can be grouped by 1-month or 3-month intervals. The number of cases in each sample will vary, but this is accounted for in the calculations of the statistic and the control limits.

Figure 5 shows a RAP chart of the QPDH dataset. This reveals that, although the observed mortality rate dropped over the 3 years of analysis, there were still episodes when it approached or exceeded the upper 95% control limit of the predicted mortality rate.

Observed minus expected charts: VLAD charts

The VLAD (variable life-adjusted display) shows the difference between the observed and the predicted deaths plotted against the patient sequential number. The SMR is related to the slope of the plot: a horizontal plot indicates an SMR of about 1. For a period where the plot rises, there are more deaths than expected, and where the plot goes down, there are fewer than expected. However, this plot (and other CUSUM-type graphs) do not give information about the severity of illness.

Figure 6 is a VLAD chart of the QPDH data. It shows, with the upward slope, those periods during the analysis when the observed deaths were higher than those predicted by the APACHE III model.

Resetting RASPRT

The RACUSUM is closely related to the RASPRT (risk-adjusted sequential probability ratio test) chart, which will be described first.

The SPRT chart uses a sequential sampling technique to test one hypothesis against another hypothesis. When the accumulating SPRT statistic reaches a critical level, the hypothesis is accepted; when it falls below a critical level, a null hypothesis is accepted as more likely. The resetting occurs when these decision barriers are crossed, and monitoring is recommenced. An RASPRT incorporates the predicted risk of death into the calculation of the SPRT statistic. This chart is used in the United Kingdom Intensive Care National Audit and Research Centre (ICNARC) casemix program reports,[6] and other applications have been reviewed.[23,24]

In a critical care context, for example, we may examine the proposal that the risk of death is increased. The null hypothesis is that the probability of death is correctly estimated by the risk-adjustment tool, while the alternative hypothesis is that mortality rate is double the odds of death compared with the probability estimated by a risk-adjustment tool such as APACHE III-J.

As each patient is discharged, the estimated probability of death and the actual outcome are compared. The statistic accumulates evidence for or against the hypotheses, based on the patient's observed outcome, the predicted risk of death, and the level of alternative performance defined by the hypothesis. Evidence accumulates to support the doubling of odds of death hypothesis over the null hypothesis when the plot moves upwards, and to support the null hypothesis over the doubling of odds of death hypothesis when the plot moves downwards.

The upper decision thresholds provide guides to the strength of the evidence (ie, the level of certainty provided by the *P* value for that threshold). A conventional arrangement has the two thresholds at a *P* of about 0.05 and a *P* of about 0.01, respectively, to indicate increasing probability that (for example) the odds of death are twice that of the risk-adjustment model predictions. Accepting the doubling of odds of death hypothesis means that there is a statistically significant trend in the data. We are advised to support that the odds of death have doubled rather than to accept that the odds of death are as predicted by the risk-adjustment model. Once the threshold is crossed, resetting to zero would occur, and appropriate steps to investigate the signal and monitoring would recommence.

Strictly, the performance of the RASPRT chart should be assessed by ARL to signal over the range of clinical scenarios. However, the use of Type I and Type II error thresholds provides a useful and familiar approximation. Care should be exercised when multiple analyses increase the risk of false positive results.

Like the RACUSUM, the RASPRT chart performances are sensitive to data errors and poor risk-adjustment model fit, and this can lead to spurious interpretation. The test is on the observed outcomes relative to the predicted outcomes. When the data cross a threshold, we need to be cautious about overinterpreting the results. An alarm signal means there may be a trend away from the benchmark. Further analysis and monitoring are required to confirm this provisional finding. Any errors or gaps in the data will distort the analysis and may bias the results, particularly when the population size is small.

There are also decision thresholds below the zero line representing the certainty with which the null hypothesis can be accepted over the primary hypothesis of (say) doubling of the odds of death. So a RASPRT statistic that crosses the lower threshold ($P \sim 0.05$ or 0.01) provides evidence that the primary hypothesis should be rejected. When the doubling of odds of death hypothesis is rejected in favour of the null hypothesis, the chart is reset and monitoring recommences.

The thresholds taken into account by the RASPRT chart design are the changes in performance that are to be detected (the odds ratio of interest), and the levels of certainty (*P* value) required to accept or reject the hypothesis. The calculation of the statistic and thresholds are shown in the Appendix. The choice of hypothesis to test (either increased or decreased odds of death) and the level of proof required are clinical or management decisions. A chart that monitors for doubling of odds of death is specifically monitoring for runs of increased mortality, while a chart that tests a halving of odds of death is monitoring for runs of improved survival.

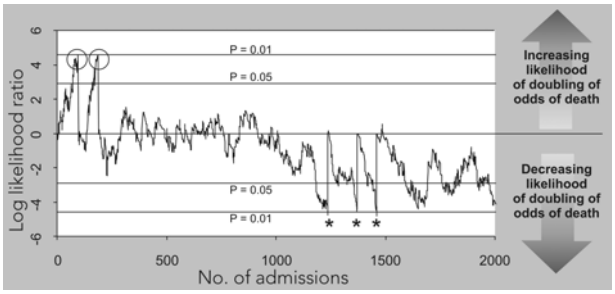
Where the statistic remains between the two sets of thresholds, the chart continues to gather evidence. It is important that pre-hoc thresholds are imposed, and that the charts reset when the threshold is reached to prevent accumulation of excessive credit for good (or bad) performance, which might delay detection of changes in the opposite direction.

Figure 7 is an example of a resetting RASPRT chart from the QPDH, where the hypothesis of doubling odds of death compared with APACHE III-J is initially accepted and then, after resetting on two occasions, is rejected in favour of the hypothesis that the APACHE III-J model is a more accurate estimate of probability of death.

Risk-adjusted CUSUM

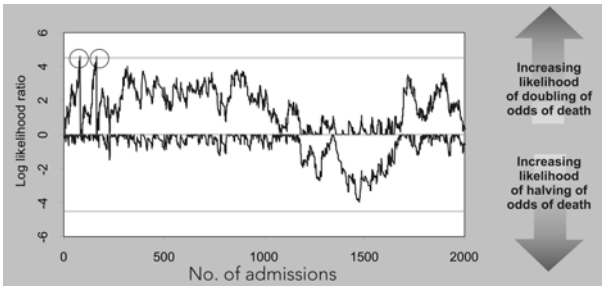
A variation of the RASPRT chart is the RACUSUM. Like the RASPRT, the RACUSUM accumulates evidence of the performance of the process, and signals when a decision threshold is crossed.

Figure 7. Risk-adjusted sequential probability ratio test (RASPRT) chart for QPDH, 2000–2003



A crossing of the uppermost boundary (circled) indicates a doubling of the odds of death compared to that predicted by APACHE III. A crossing of the lowermost boundary (starred) indicates there is not a doubling of the odds of death compared to that predicted by APACHE III. When these boundaries are crossed, there is a 1% chance that this finding is due to chance.

Figure 8. Risk-adjusted cumulative sum (RACUSUM) chart for QPDH, 2000–2003



A crossing of the uppermost boundary (circled) indicates a doubling of the odds of death compared to that predicted by APACHE III. When this boundary is crossed, there is a 1% chance that the finding is due to chance.

In contrast to the RASPRT, the RACUSUM[21,22] imposes a lower absorbing barrier at the zero line. The analysis can monitor for an increased probability of death, say doubling of the odds of death compared with the predictions provided by the risk-adjustment model. However, there is no opportunity to accumulate "credit" for good performance as the RACUSUM statistic is bounded by the lower limit of zero. The charts can be paired to monitor simultaneously for increased and decreased patient mortality, for example testing for runs of doubling and halving of odds of death.

Figure 8 shows a combined chart that monitors both doubling and halving of the odds of death compared with APACHE III-J at the QPDH. The same trends are reliably reproduced as in the previous charts. Note that while the previous RASPRT accepts the likelihood of the adequacy of fit of the APACHE III model, the two RACUSUMs are only testing for a doubling or halving of the odds of death. Hence, both the RASPRT chart and the RACUSUM signal early during the analysis the high probability of doubling of odds of death. However, the RASPRT signals later in the analysis, supporting acceptance of the APACHE model as the better option. The lower RACUSUM testing for halving of odds of death does not signal, as there are no runs supporting a halving of the odds of death.

The RACUSUM was proposed for adult critical care unit monitoring,[8] and is used as part of the reports provided by the Australian and New Zealand Paediatric Intensive Care registry since October 2004.[30]

Risk-adjusted EWMA

The EWMA (exponentially weighted moving average) is a running estimate of the mean output of a process, where the most recent observations are given exponentially more weight than historically distant observations. It is an excellent estimate of the current mean of a process where the mean is changing slowly.

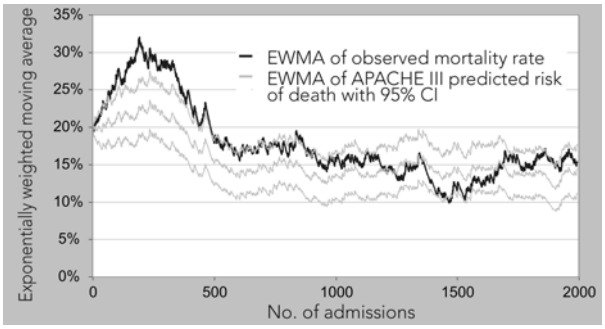
Risk adjustment is introduced as the RAEWMA compares the weighted mean of the observed mortality rate to a similarly weighted mean average of the expected outcomes, or weighted mean predicted mortality rate. Confidence intervals are calculated around the predicted mortality rate. The chart provides a running comparison of expected and observed outcomes, weighting the most recent events, and allowing historical observations to become exponentially less influential.

The RAEWMA chart has equivalent performance to the RACUSUMs in detecting small changes.[31] It has a dependable Gaussian distribution and is suited to analysis of a stream of individual observations, such as patient outcomes and probability estimates.[32] There is an inherent smoothing function, so the RAEWMA is robust to chance variations resulting from poor model discrimination, as long as the calibration of the model is accurate, thus smoothing out random fluctuations in outcomes. The direct comparison of the predicted and observed mortality rates over time is relatively easy to explain. However, if large changes in mortality performance are possible, the weight must be increased, or alternatively the RAEWMA should be combined with SMRs or *p*-charts to provide more rapid detection.

Figure 9 shows the, by now familiar, hypothetical QPDH dataset, with an excess mortality initially well above the expected mortality rate with APACHE III-J, which approaches the expected rate during the analysis.

Note that in this example, the chart was not reset when the observed mortality rate fell outside the control limits. Thus, this chart shows a running comparison between the weighted estimates of the current risk of death predicted by

Figure 9. Exponentially weighted moving average (EWMA) chart for QPDH, 2000–2003



EWMA of observed mortality is plotted with the EWMA for APACHE III predicted risk of death. When observed mortality varies above the 95% CI for APACHE III predicted risk of death, there is a less than 2.5% probability that this finding is due to chance.

the model with confidence intervals, against the current weighted estimate of the mortality rate among the same patients.

The chart can be modified to signal and be reset, like the VLAD, RACUSUM and RASPRT charts, or to provide a running estimate of the SMR and the normalised discrepancy between observed and expected values.

Governance and clinical monitoring

Monitoring clinical performance is complex and resource-intensive. Choice of a robust risk-adjustment model, an appropriate method for data analysis, and the preferred chart or display format are important decisions.

When outcome monitoring is undertaken, there is a responsibility to act in a considered, timely and appropriate manner to address whether there is a data and model-related anomaly, or a clinical issue. A monitoring and response strategy using the VLAD has been introduced in Queensland hospitals.[33]

It is essential to develop a plan for follow-up and analysis of the circumstances surrounding any statistical anomalies that are detected. Part of this is to examine the quality and completeness of data collection and the adequacy of model fit, to exclude a spurious cause for any detected changes in performance. The ANZICS Core Management Committee is developing appropriate responses to the statistical anomalies that will inevitably be detected, in consultation with the constituent state and territory health authorities. If data- and model-related anomalies cannot be excluded, there must be consideration of the possibility of a true difference in clinical performance.

Misuse of data and misinterpretation of results may lead to serious and avoidable consequences, such as falsely classifying an "outlier" hospital. Any process for the widespread implementation of clinical monitoring, such as that contemplated by national (ANZICS) and state (eg, Queensland Health and VICDRC) data committees, requires appropriate and transparent guidelines. This must include a transparent consensus process for identifying and supporting apparent "outliers". These issues are beyond the scope of this review.

Summary

Risk-adjusted analysis of mortality outcomes seeks to control for variation in mortality rates caused by differences in severity of illness, physiological reserve, diagnosis and casemix. By adjusting for the factors that can be statistically modelled, it is possible to use risk-adjusted methods to screen for statistical anomalies across populations, or over a period in a single context. Monitoring is part of a coordinated risk and quality measurement strategy, complementing measurement of key performance indicators and investigations of index events.

The most important sources of error arise from poor quality data, poor quality risk-adjustment models, insufficient population size, and misinterpretation of a correctly constructed and accurate analysis.

At present, risk-adjusted analysis is best viewed as a screening procedure. If anomalies are identified, a review of data quality and completeness, as well as data analysis and other operation issues should proceed. Often the aberrations are explained, but changes in the quality of the process of care can manifest as improved or deteriorating measurements of risk-adjusted clinical outcome.

Author details

David A Cook, Staff Specialist[1,2]
Graeme Duke, Director, Critical Care Department[1,3]
Graeme K Hart, Chair[1,4,5]
David Pilcher, Staff Specialist[1,6]
Daniel Mullany, Director[1,7]
1 Australian and New Zealand Intensive Care Society CORE Management Committee, Melbourne, VIC.
2 Intensive Care Unit, Princess Alexandra Hospital, Brisbane, QLD.
3 Critical Care Department, Northern Hospital, Melbourne, VIC.
4 Department of Intensive Care and Austin Centre for Applied Clinical Informatics, Austin Hospital, Melbourne, VIC.
5 Intensive Care Unit, Warringal Private Hospital, Melbourne, VIC.
6 Intensive Care Unit, Alfred Hospital, Melbourne, VIC.
7 Intensive Care Unit, Prince Charles Hospital, Brisbane, QLD.
Correspondence: d.cook@mailbox.uq.edu.au

References

1 Tunnell RD, Millar BW, Smith GB. The effect of lead time on severity scoring, mortality prediction and SMR in intensive care. *Anaesthesia* 1998; 53: 1045-53.

2 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.

3 Knaus WA, Draper EA, Wagner DP, et al. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818-29.

4 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: 1619-36.

5 Harrison DA, Parry GJ, Carpenter JR, et al. A new risk prediction model for critical care: the Intensive Care National Audit and Research Centre (ICNARC) model. *Crit Care Med* 2007; 35: 1091-8.

6 Intensive Care National Audit and Research Centre (ICNARC). A guide to the ICNARC Case Mix Programme. Data Analysis Report. London, 2007. Available at: www.ICNARC.org (accesed May 2008).

7 Cockings JGL, Cook DA, Iqubal RK. Risk adjusted process monitoring in intensive care using cumulative expected minus observed mortality. *Crit Care* 2006; 10: R28.

8 Cook D, Steiner S, Cook R, et al. Monitoring the evolutionary process of quality: risk-adjusted charting to track outcomes in intensive care. *Crit Care Med* 2003; 31: 1676-82.

9 Cook DA. Methods to assess performance of models estimating risk of death in ICU patients: a review. *Anaesth Intensive Care* 2006; 34: 164-75.

10 Cook D, Joyce C, Barnett R, et al. Prospective independent validation of APACHE III models in an Australian tertiary adult intensive care unit. *Anaesth Intensive Care* 2002; 30: 308-15.

11 Cook DA. Performance of APACHE III models in an Australian ICU. *Chest* 2000; 118: 1732-8.

12 Kramer AA. Predictive mortality models are not like fine wine. *Crit Care* 2006; 9: 636-7.

13 Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. *Crit Care Med* 2006; 34: 1297-310.

14 Woodall WH. The use of control charts in health-care and public-health surveillance. *J Qual Technol* 2006; 38: 89-104.

15 Grigg O, Farewell VT. An overview of risk adjusted charts. *J R Statist Soc A* 2004; 167: 523-39.

16 Hart MK, Lee KY, Hart RF, et al. Application of attribute control charts to risk adjusted data for monitoring and improving health performance. *Qual Manag Health Care* 2003; 12: 5-19.

17 Cook DA. The development of risk adjusted control charts and machine learning models to monitor the mortality rate of intensive care unit patients. Brisbane: University of Queensland, 2004.

18 Lovegrove J, Valencia O, Treasure T, et al. Monitoring the results of cardiac surgery by variable life-adjusted display. *Lancet* 1997; 350: 1128-30.

19 Sherlaw-Johnson C. A method for detecting runs of good and bad clinical outcomes on variable life-adjusted display (VLAD) charts. *Health Care Management Science* 2005; 8: 61-5.

20 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: 1697-700.

21 Steiner S, Cook R, Farewell V. Risk adjusted monitoring of surgical outcomes. *Med Decis Making* 2001; 21: 163-9.

22 Steiner S, Cook R, Farewell V, et al. Monitoring surgical performance using risk-adjusted cumulative sum charts. *Biostatistics* 2000; 1: 441-52.

23 Grigg OA, Farewell VT, Spiegelhalter DJ. Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts. *Stat Methods Med Res* 2003; 12: 147-70.

24 Spiegelhalter D, Grigg O, Kinsman R, et al. Risk adjusted sequential probability ratio tests: applications to Bristol, Shipmann and adult cardiac surgery. *Int J Qual Health Care* 2003; 15: 7-13.

25 Grigg OA, Spiegelhalter D. A simple risk-adjusted exponentially weighted moving average. *J Am Stat Assoc* 2007; 102: 140-52.

26 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005; 24: 1185-202.

27 Spiegelhalter D. Handling overdispersion of performance indicators. *Qual Saf Health Care* 2005; 14: 347-51.

28 Alemi F, Sullivan T. Tutorial on risk adjusted x-bar charts: applications to measurement of diabetes control. *Qual Manag Health Care* 2001; 9: 57-65.

29 Alemi F, Oliver D. Tutorial on risk adjusted *p* charts. *Qual Manag Health Care* 2001; 10: 1-9.

30 Australian and New Zealand Paediatric Intensive Care Group. Report of the ANZPICG for 2003. ANZPICG, 2004.

31 Grigg O, Spiegelhalter D. Discussion. *J Qual Tech* 2006; 38: 124-6.

32 Montgomery D. Chapter 7: Cusum and EWMA control charts. In: Introduction to statistical quality control. 3rd ed. New York: John Wiley and Sons, 1996: 313-47.

33 Duckett SJ, Coorey M, Sketcher-Baker K. Identifying variations in quality of care in Queensland hospitals. *Med J Aust* 2007; 187: 571-5.

34 Armitage P, Berry P. Inferences from proportions. In: Statistical methods in medical research. 3rd ed. London: Blackwell Scientific Publications, 1994: 118-24.

35 Vollset S. Confidence intervals for a binomial proportion. *Stat Med* 1993; 12: 809-24.

36 Sherlaw-Johnson C, Gallivan S. Approximating prediction intervals for use in variable life adjusted displays. Technical note. London: Clinical Operational Research Unit, Department of Mathematics, University College, 2000. Report No. 563. ❑

Appendix. Calculations and formulas

Standardised mortality ratio (SMR)

There are n patients indexed by i. $\hat{\pi}_i$ is the estimate of the probability of death provided by the model. For a patient who dies, the outcome is $Y_i = 1$; for a patient who survives, it is $Y_i = 0$.

$$SMR = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} \hat{\pi}_i}$$

Confidence intervals estimate the precision of the SMR based on various assumptions about the relevant distributions. Confidence intervals around the numerator assume a binomial distribution, which is defined by sample size and observed deaths.

Approximations of the binomial distribution have been reviewed,[34,35] and the choice is a balance between simplicity and accuracy. The examples in Figures 1 and 2 use a continuity-corrected quadratic method. The ANZICS CORE reports use an exact method based on the F distribution. There is probably little difference given the size of the samples, the SMR ranges and the inaccuracies inherent in the data and the models.

Caterpillar plot

SMR and confidence intervals for each provider are calculated, and the providers are ranked according to SMR. The rank (y axis) is plotted against the SMR.

Funnel plot

The approach described by Spiegelhalter[26] describes setting a standardised mortality rate based on the average mortality rate across the pooled samples of patients. This has been adapted to show an SMR, to make it easier to describe and interpret, rather than a nominal average death rate, although the methods are essentially the same.

The SMR for each unit is calculated, as described above. The number of cases in each provider's sample is used to plot the x coordinate, and the y axis shows the SMR. It is customary to identify one or more of the units or providers of interest.

The control limits around an SMR of 1 are plotted to emphasise that the larger the sample, the more certain the observed mortality rate. Control limits are estimated using approximations to the binomial distribution, with the sample size and the average mortality rate providing the number of cases and the number of deaths, from which is estimated the distribution. Several methods can be used to characterise the possible distribution. Given the possible sources of error, a simple normal approximation is used in the example in Figure 4.

Risk-adjusted p (RAP) chart

The following additional notation is used for RAP charts. The samples are indexed by i, and within the samples the cases are indexed by j. The expectation of the event Y is

$$\hat{E}(Y_{ij}) = \hat{\pi}_{ij}$$

and the variance is estimated as

$$\hat{\mathrm{var}}(Y_{ij}) = \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})$$

with the observed and predicted mortality rates

$$R_i = \frac{\sum Y_{ij}}{n_i} \quad \hat{E}(R_i) = \frac{\sum \hat{\pi}_{ij}}{n_i}$$

The variance of the observed mortality rate is

$$\hat{\mathrm{var}}(R_i) = \frac{\sum_{j=1}^{n_i} \hat{\mathrm{var}}(Y_{ij})}{n_i^2} = \frac{\sum_{j=1}^{n_i} \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}{n_i^2}$$

The RAP chart compares R_i to $\hat{E}(R_i)$ with control limits calculated around $\hat{E}(R_i)$ and defined as multiples of the standard deviation

$$CL_i = \frac{\sum \hat{\pi}_{ij}}{n_i} \pm a.\sqrt{\frac{\sum_{j=1}^{n_i} \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}{n_i^2}}$$

The chart plots R_i, $\hat{E}(R_i)$ and control limits against the index i, which often conveniently maps onto a month or other time period for the sample.

Observed minus expected charts: VLAD

The equation for calculating the statistic is

$$C_n = \sum_{i=1}^{n} (\hat{\pi}_j - Y_j)$$

The Sherlaw–Johnson VLAD (variable life-adjusted display)[19,36] calculates control limits around the VLAD statistic, based on an equivalent performance to the RACUSUM (risk-adjusted cumulative sum) under the same conditions. In fact, the limits are reset at the times when an equivalent CUSUM "alarms" and is reset. The advantage of this technique is that it allows trends to be

more easily appreciated qualitatively than with the RACUSUM, and it is equivalent in performance to both the RACUSUM and the RAEWMA (risk-adjusted exponentially weighted moving average). This is the method adopted by Queensland Health[33] to display risk-adjusted surveillance data.

RASPRT and RACUSUM

Both the resetting RASPRT (risk-adjusted sequential probability ratio test) and the RACUSUM are applications of the sequential probability ratio test of Wald from the 1940s.

The RACUSUM method provides a test of the hypothesis that the risk-adjusted odds of death are unchanged, against an alternative hypothesis of a changed odds ratio (OR). An upper RACUSUM with control limits tests the null hypothesis of unchanged odds of death; the alternative hypothesis is that the OR has increased. A lower RACUSUM with control limits tests the null hypothesis of unchanged odds of death; the alternative hypothesis is that the OR has decreased. In the example in Figure 8, the RACUSUM procedure is designed to best detect a doubling (or halving) of the OR.

A score (w_j) is given to each patient. It is derived from the log-likelihood ratio of the current risk of death, compared with the risk of death if the overall level of ICU performance has changed. Under the null hypothesis, the likelihood for patient j is given by $\pi_j^{1-Y_j}(1-\pi_j)^{Y_j}$ and the odds of death are

$$\frac{\pi_j}{(1-\pi_j)}.$$

Under the alternative hypothesis, the odds of death are

$$\frac{OR_A\pi_j}{1-\pi_j}.$$

As there are only two possible outcomes (death or survival), the two possible log-likelihood ratio scores for patient *j* are given by

$$w_j = \log\left[\frac{OR_A}{(1-\pi_j + OR_A\pi_j)}\right]$$

if the patient dies, or

$$w_j = \log\left[\frac{1}{(1-\pi_j + OR_A\pi_j)}\right]$$

if the patient survives.

For the upper control chart, an upper CUSUM statistic, S_j^+ is plotted against *j*, where *j* is the patient number and $S_0^+ = 0$.

$$S_j^+ = \max\left(S_{j-1}^+ + w_j, 0\right)$$

The RACUSUM formally tests the null hypothesis, H_0: $OR_0 = 1$, against the alternative hypothesis, H_A: $OR_A > 1$. Successive non-negative values lead to accumulation of S_j^+ until its value exceeds the control limit, h^+.

For testing for a change in the risk-adjusted odds of death where the mortality is falling, the procedure is similar to the test for an increased OR. For the convenience of plotting both charts on the same figure, the statistic S^- is accumulated as a negative value (or zero) and h^- is a negative value.

$$S_j^- = \min\left(S_{j-1}^- - w_j, 0\right)$$

The OR_A is less than 1 and the control limit is h^-, the value below which S^- must fall to give an alert or alarm.

The RACUSUM is efficient in detecting small changes in the process, if properly constructed, and is identical to the VLAD (as described) and equivalent to the RAEWMA. It does not build up credit for past performance, due to the absorbing barrier at the zero line, and will tend to detect changes faster than the RASPRT. However, it is very difficult to detect qualitative trends from examining the chart and it can be difficult to explain.

The RASPRT procedure is very similar, although the hypotheses to be tested differ subtly. The null hypothesis is usually that the risk of death is accurately estimated by the risk-adjustment tool, but the alternative hypothesis is that the probability of death is more accurately predicted by a change in the odds of death, defined by an OR.

The RASPRT statistic after patient *j* is

$$R_j = R_{j-1} + w_j$$

where the weight, w_j to be accumulated is defined the same way.

However, the decision thresholds are *a* and *b*. At or above *a*, the probability of eventually incorrectly rejecting the null hypothesis in favour of the hypothesis that the odds of death are truly doubled (Type I error) occurs with a probability of α, which can be set at (say) 0.05 or 0.01. Whereas, at or below *b* the probability of eventually incorrectly rejecting the hypothesis that the odds of death are doubled, in favour of the null hypothesis that the

probability of death is accurately estimated by the risk-adjustment tool (Type II error) occurs with a probability of β, which can be set at (say) 0.05 or 0.01.

$$a = \log\left(\frac{\beta}{1-\alpha}\right)$$

$$b = \log\left(\frac{1-\beta}{\alpha}\right)$$

When the RASPRT is reset after reaching the critical defined levels of evidence, it is termed a resetting RASPRT. The advantages of the RASPRT are that it displays trends in performance similarly to the VLAD. Most importantly, it can be more readily analysed in terms of an approximation of Type I and Type II error, although strictly ARL analysis is the most accurate. A limitation of the RASPRT is that it is used to test the better choice between an alternative hypothesis and null hypothesis pair. For monitoring across groups of institutions, the greatest concern is about detecting increased mortality events, so this is only a problem if detecting better than predicted performance is important.

Risk-adjusted EWMA

The EWMA statistic of the observed deaths is calculated from the series of observations.

$$EWMA_j^Y = Y_j\lambda + EWMA_{j-1}^Y(1-\lambda)$$

or

$$EMWA_j^Y = (1-\lambda)^j EMWA_0^Y + \lambda\sum_{k=1}^{j}(1-\lambda)^{j-k}Y_k$$

The value of the statistic $EWMA_j^Y$ is compared with an estimate of the expected value of the EWMA statistic, using the series of $\hat{\pi}_j$,

$$\hat{E}\left(EWMA_j^Y\right) = EWMA_j^{\hat{\pi}}$$

which is calculated from

$$EWMA_j^{\hat{\pi}} = \hat{\pi}_j\lambda + EWMA_{j-1}^{\hat{\pi}}(1-\lambda)$$

or

$$EWMA_j^{\hat{\pi}} = (1-\lambda)^i EMWA_0^{\hat{\pi}} + \lambda\sum_{k=1}^{j}(1-\lambda)^{j-k}\hat{\pi}_k$$

The control limits are calculated from an estimate of the variance of $EWMA_j^Y$, assuming that the starting estimate of the mortality rate $EWMA_0^Y$ has a variance of zero.

$$\mathrm{var}\left(EWMA_j^{\hat{\pi}}\right) = \lambda^2\sum_{k=1}^{j}(1-\lambda)^{2(j-k)}\,\mathrm{var}\,Y_k$$

The control limits can be estimated using the variance of Y_j as $\hat{\pi}_j(1-\hat{\pi}_j)$:

$$CL_j = EWMA_j^{\hat{\pi}} \pm a\lambda\sqrt{\sum_{k=1}^{j}(1-\lambda)^{2(j-k)}\hat{\pi}_j(1-\hat{\pi}_j)}$$

The RAEWMA is relatively simple to program, despite the formidable formulas, and readily shows trends in mortality outcomes. The EWMA method detects small and gradual changes efficiently, and its performance is equivalent to the RACUSUM and the related VLAD with control limits. It offers considerable advantages in the presence of a smoothing factor, which can be useful in handling binary outcome data. This smoothing factor can present a delay in detection compared with a RAP chart or SMR data when changes in performance are very large.